

Accurate detection of tumor clonality and ongoing expansion mode from genomic data

Yanjie Chen^{1,2}, Roman Jaksik⁴, Peter Terranova^{1,5}, Sara El Baghdadi^{1,6,7}, Andrew Koval^{8,9},
Monika K. Kurpas⁴, Simon Tavaré^{1,2,3}, Marek Kimmel^{4,8}, and Khanh N. Dinh^{1,2}

¹Irving Institute for Cancer Dynamics, Columbia University, New York, NY, USA

²Department of Statistics, Columbia University, New York, NY, USA

³Department of Biological Sciences, Columbia University, New York, NY, USA

⁴Department of Systems Biology and Engineering, Silesian University of Technology, Gliwice,
Poland

⁵Fordham University, New York, NY, USA

⁶École Polytechnique, Paris, France

⁷Institute for Computational and Mathematical Engineering, Stanford University, Stanford, CA,
USA

⁸Departments of Statistics and Bioengineering, Rice University, Houston, TX, USA

⁹Benaroya Research Institute, Seattle, WA, USA

June 15, 2026

Abstract

Recent evidence shows that despite considerable effort, currently available algorithms for estimating intra-tumor heterogeneity (ITH) remain limited. We developed DECODE (Deciphering Cancer Origin from DNA Evolution), a novel mutation clustering method that incorporates the impact of sample-specific sequencing coverage and mutation calling biases. On synthetic data, DECODE outperformed existing methods across multiple clonality metrics and accurately detected and characterized the neutral tail in the site frequency spectrum (SFS), which encodes the tumor’s ongoing expansion mode. In acute myeloid leukemia, accounting for the neutral tail enabled DECODE to yield more parsimonious clonal decompositions that align more closely with known subclonal dynamics that drive relapse. Applied to data from The Cancer Genome Atlas, DECODE not only detected a neutral SFS tail in most samples across tumor types but also uncovered a clinically meaningful link between ITH and survival in low-grade glioma. By jointly inferring clonality and expansion mode, DECODE provides two complementary and prognostically relevant readouts of tumor evolution from single tumor genomic samples.

1 Introduction

Intra-tumor heterogeneity (ITH) has been consistently associated with worse cancer patient outcome [Andor et al., 2016], drug resistance [McGranahan and Swanton, 2017, Dagogo-Jack and Shaw, 2018], and increased risk of relapse and metastasis across different cancer types [Turajlic and Swanton, 2016, Caswell and Swanton, 2017]. Subclonal reconstruction has accordingly become a pillar of cancer evolution research, with growing applications in patient stratification and treatment design [Tarabichi et al., 2021]. Many such methods operate on the site frequency spectrum (SFS), the distribution of somatic mutations by their variant allele frequency (VAF) within a tumor. The shape of the SFS encodes evolutionary history. Clonal mutations form a cluster at high VAFs, and each subclone contributes a lower-VAF cluster. Tumor deconvolution decomposes this spectrum to recover the number and sizes of subclones present. However, despite the proliferation of subclonal reconstruction algorithms, recent studies have exposed substantial limitations that remain unaddressed. The ICGC-TCGA DREAM Somatic Mutation Calling Tumor Heterogeneity and Evolution Challenge found that, among 31 single-sample mutation decomposition algorithms, no single method achieved top performance across all aspects of clonal reconstruction [Salcedo et al., 2025]. Notably, the authors found that algorithm choice influenced performance more strongly than any tumor feature, including purity, copy-number state, and sequencing depth. In addition, ensemble strategies were unable to outperform the best individual methods. Together, these results underline an urgent need for more accurate and robust methods to estimate ITH from cancer genomic data.

We hypothesized that addressing two often-overlooked factors can improve subclonal reconstruction. First, regardless of whether the tumor evolves neutrally or undergoes selective sweeps, neutral mutations that accumulate as cells divide generate a characteristic power-law tail in the SFS [Williams et al., 2016]. The exponent of the SFS tail distribution quantifies the ongoing rate of mutation accumulation relative to cell division, and thus reflects the tumor’s expansion dynamics. Population genetics theory shows that the tail power takes the value 1 in constant-size populations [Watterson, 1975] and 2 under exponential growth [Griffiths and Tavaré, 1998]. To our knowledge, MOBSTER currently remains the only method that explicitly models the neutral tail [Caravagna et al., 2020]. Indeed, the DREAM Challenge found that applying MOBSTER to remove tail mutations before deconvolution reduces the misassignment of neutral variants to spurious subclones [Salcedo et al., 2025]. However, MOBSTER’s authors noted that the algorithm requires high sample purity and coverage of $\geq 100\times$ to detect the tail reliably, exceeding the depth of most available datasets [Caravagna et al., 2020]. The second critical but frequently neglected factor is that sample-specific features can strongly distort the SFS. We previously showed

that sequencing coverage, particularly when low or non-uniform as typical of real data, strongly shapes both the dispersion of mutation clusters and the form of the observed SFS tail [Dinh et al., 2020]. Additionally, mutation callers typically discard a large fraction of mutations at low frequencies, which disproportionately depletes and distorts the empirical SFS tail [Krøigård et al., 2016]. Failure to account for these biases may explain why many current methods infer clonality inaccurately and why MOBSTER cannot reliably detect the tail in low-coverage data.

2 Results

2.1 DECODE incorporates corrections for sequencing coverage and variant calling to improve detection of mutation clusters and neutral tails

We developed DECODE (Deciphering Cancer Oorigin from DNA Evolution), a novel mutation clustering algorithm based on our previous mathematical results in [Dinh et al., 2020]. DECODE assumes that a sample’s exact site frequency spectrum (SFS) consists of H clusters, formed of mutations present in the most recent common ancestors (MRCAs) of each subclone, and a neutral tail containing other mutations that arise as cells of all subclones divide (Fig. 1A). The expected number of mutations present in k cells in a sample of n cells thus follows

$$\mathbb{E}(S^n(k)|\theta_H) \approx \omega_0 \cdot g_{\text{tail}}(k|\alpha) + \sum_{h=1}^H \omega_h \cdot g_{\text{peak}}(k|p_h)$$

where the ω_0 mutations in the tail follow the power law $g_{\text{tail}}(k|\alpha) \propto 1/k^\alpha$ and the ω_h mutations in each cluster h follow the binomial distribution $g_{\text{peak}}(k|p_h)$ with mean variant allele frequency (VAF) p_h . Given sequencing coverage φ , where $\varphi(r)$ is the probability that a genomic site is sequenced with a total of r reads, we have shown that the expected mutation count with observed VAF in a bin $(f_1, f_2] \subset [0, 1]$ follows [Dinh et al., 2020]

$$\mathbb{E}(S((f_1, f_2])|\theta_H) = \sum_{k=1}^n \mathbb{E}(S^n(k)|\theta_H) \sum_r \varphi(r) \cdot \mathbb{P}\left(\text{Binomial}\left(r, \frac{k}{n}\right) \in (f_1 r, f_2 r]\right)$$

The formula can be further modified to accommodate data filtering, which is a critical step in mutation calling and tends to predominantly remove mutations with few variant reads or low coverage in order to reduce the rate of false positives [Krøigård et al., 2016]. Given thresholds (L, M) where mutations with fewer than L variant reads or M total reads are removed, the expected mutation count in bin $(f_1, f_2]$ becomes

$$\mathbb{E}\left(S^{L,M}((f_1, f_2])|\theta_H\right) = \sum_{k=1}^n \mathbb{E}(S^n(k)|\theta_H) \sum_r \varphi_{L,M}(r) \cdot \mathbb{P}\left(\text{Binomial}\left(r, \frac{k}{n}\right) \in (f_1 r, f_2 r] \cap [L, \infty)\right) \quad (1)$$

where the coverage distribution $\varphi_{L,M}(r)$ is computed from the retained mutations. We observed that thresholds L and M impact the SFS in different ways [Dinh et al., 2020]. Removing mutations with low support (i.e., increasing L) mainly reduces the tail, as it is composed of mutations with the lowest VAFs (Fig. 1A). Meanwhile, filtering based on coverage (e.g., by altering M) eliminates mutations across the SFS uniformly, and thus only reduces the scale of the SFS without significantly changing its shape.

DECODE takes advantage of our mathematical framework to seek the parsimonious SFS decomposition that is consistent across different data filtering routines. Given a DNA-sequencing sample, it first selects thresholds (L, M) for three different data subsamples, termed *inference A*, *inference B* and *validation* (**step 1**, Fig. 1B). The variant read count threshold is different in each subset, so the SFS from each filtered subsample assumes a different shape (**step 2**, Fig. 1B).

For a given cluster count H , DECODE infers the parameters $\theta_H = (\alpha, p_1, \dots, p_H, \omega_0, \omega_1, \dots, \omega_H)$, which characterize the exponent of the tail, each cluster’s mean VAF, and the mutation count in each component. The estimation is obtained via ABC-SMC-DRF (Approximate Bayesian Computation sequential Monte Carlo with distributional random forest) [Dinh et al., 2025], a general likelihood-free inference method that we recently introduced. ABC-SMC-DRF is modified to maintain cluster ordering, where cluster 1 is truncal and cluster H corresponds to the rarest subclone. DECODE thus finds the distribution of θ_H such that the SFS predicted by Eq. 1 best matches the empirical SFS from *inference A* and *inference B* subsamples (**step 3**, Fig. 1B).

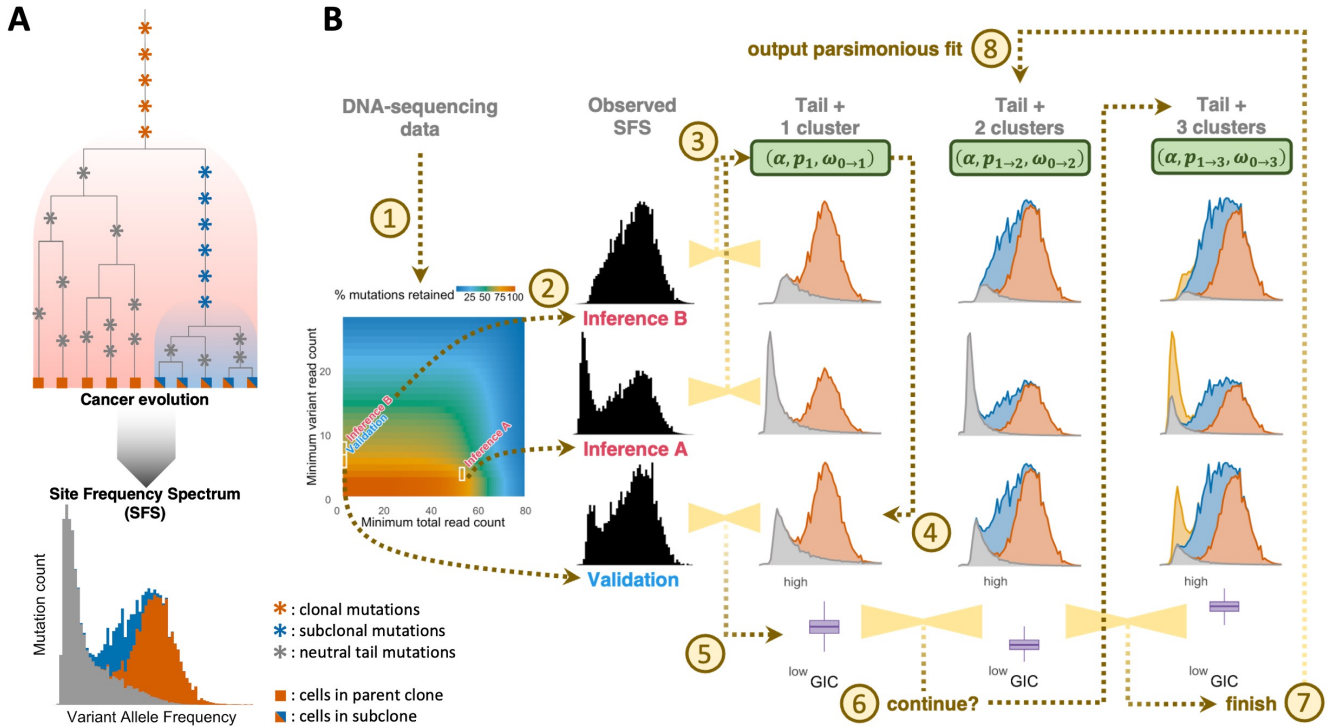


Figure 1: **A**: DECODE’s model for clonal evolution and the corresponding Site Frequency Spectrum (SFS). Clonal mutations are present in a tumor’s MRCA and shared across all cancer cells, forming the truncal cluster in the SFS. Each subclone results from an ongoing selection sweep, where the mutations in the subclone’s MRCA are present in its cells and constitute a subclonal cluster at lower Variant Allele Frequencies (VAFs). Mitoses of cells within all subclones induce additional mutations, resulting in a neutral tail in the lowest-VAF region of the SFS. **B**: Schematic overview of DECODE’s algorithm. For each genomic sample, DECODE extracts the SFS from three filtering strategies, based on variant and total read counts (**step 1-2**). For each cluster count H , it estimates model parameters with ABC-SMC-DRF from two inference SFS subsamples (**step 3**), then finds the Generalized Information Criterion (GIC) density based on the predicted SFS for the validation subsample (**step 4-5**). The GIC distributions from successive H ’s are compared to find the parsimonious fit (**step 6-8**).

To determine the parsimonious cluster configuration, DECODE tests the capacity of the inferred parameter distribution to predict the SFS from the *validation* subsample (**step 4**, Fig. 1B). The accuracy of the prediction, balanced against the model’s complexity as determined by cluster count H , is quantified with the Generalized Information Criterion (GIC) [Konishi and Kitagawa, 1996] (**step 5**, Fig. 1B). DECODE compares the GIC densities based on an increasing sequence of cluster counts. It selects the more complex result as the better model if its GIC is lower, and continues adding more clusters (**step 6**, Fig. 1B). Otherwise, it selects the model with the lower cluster count (**step 7**, Fig. 1B) as the most parsimonious decomposition (**step 8**, Fig. 1B). DECODE fits the data with and without a neutral tail, then selects the better model by GIC.

Two distinctions set DECODE apart from previous methodologies. First, the incorporation of mutation filtering strategies enables it to validate its results even for single-sample data. This reduces the risk for DECODE to either overfit or underfit the mutation clusters. Second, incorporating sample-specific sequencing coverage directly in the mathematical framework both enhances DECODE’s detection of overlapping clusters and ensures unbiased estimation for neutral tail power. Many existing methods assume mixture models for the SFS where each cluster density is characterized with parameters for the mean and dispersion [Carter et al., 2012, Roth et al., 2014, Miller et al., 2014, Deshwar et al., 2015, Caravagna et al., 2020]. Their criteria for selecting the parsimonious model thus strongly penalize decompositions with significantly overlapping clusters, to prevent overfitting due to the high parameter count. In contrast, the distribution for each cluster h in DECODE is defined only with its mean VAF p_h , whereas its dispersion and overall shape are determined by the coverage distribution. We therefore expect that DECODE is more sensitive at detecting overlapping clusters, to be confirmed in the next sections. Furthermore, non-uniform coverage most significantly affects the low-VAf region in the SFS, which distorts the shape of the neutral tail [Dinh et al., 2020]. By accounting for this effect, DECODE is able to detect and characterize the tail more accurately and consistently.

2.2 DECODE outperforms existing methods in inferring tumor clonality and detecting neutral tail from the SFS

We first examined DECODE’s performance in inferring different aspects of tumor heterogeneity from single genomic samples by utilizing the ICGC-TCGA DREAM Somatic Mutation Calling Tumor Heterogeneity and Evolution Challenge. The DREAM Challenge consisted of 51 synthetic tumors, simulated under different assumptions of clonality, copy number aberrations, cellularity, sequencing depth and spatial growth that mirror real tumors [Salcedo et al., 2020]. The group compared 31 subclonal reconstruction algorithms and five reference methods based on the accuracy of the inferred sample purity (*task 1A*), subclone count (*task 1B*), cellular prevalences and mutation counts of each subclone (*task 1C*), deterministic and probabilistic assignments of mutations to subclones (*task 2A-B*), and deterministic and probabilistic subclonal phylogeny (*task 3A-B*) [Salcedo et al., 2025]. As DECODE was designed to cluster mutations with similar copy numbers, we limited the analysis to predominantly diploid samples ($n = 29$). DECODE currently does not support the prediction of subclonal phylogenies, so our comparison focused on *tasks 1A-2B*. Since DREAM concentrated on the accuracy of mutation clustering, we implemented DECODE without the tail compartment, and compared its performance against the 36 software packages benchmarked in [Salcedo et al., 2025].

DECODE ranks among the top methods for purity inference (Fig. 2A). Its mean score across samples, normalized to the range between best and worst software akin to the DREAM protocol [Salcedo et al., 2025], is 7% worse than PhyloWGS, the leading method [Deshwar et al., 2015]. DECODE’s average score is most impacted by sample S5, where the synthetic tumor contains two independent clones, both at 40% cellular prevalence (i.e., VAF 20%), yielding an overall tumor purity of 80% (Supplementary Fig. 1A). However, DECODE’s predicted purity is based on the assumption that the truncal cluster in the SFS consists of mutations present in every tumor cell (Fig. 1A). Therefore, although it correctly identifies a truncal cluster at $p_1 = 0.2$ in S5, DECODE incorrectly infers a purity of $2 \cdot p_1 = 40\%$ for sample S5, similar to other approaches that do not take copy numbers into account (Fig. 2A, Supplementary Fig. 1B, C). DECODE’s predicted purity is much closer to the true values for samples where cancer cells share a common ancestor, as we expect to be the case for many real tumors.

DECODE’s inferences for subclone count and subclonal prevalences outperform algorithms benchmarked in the DREAM Challenge (Fig. 2B, C), with 2% and 5% improvements in normalized mean scores compared to

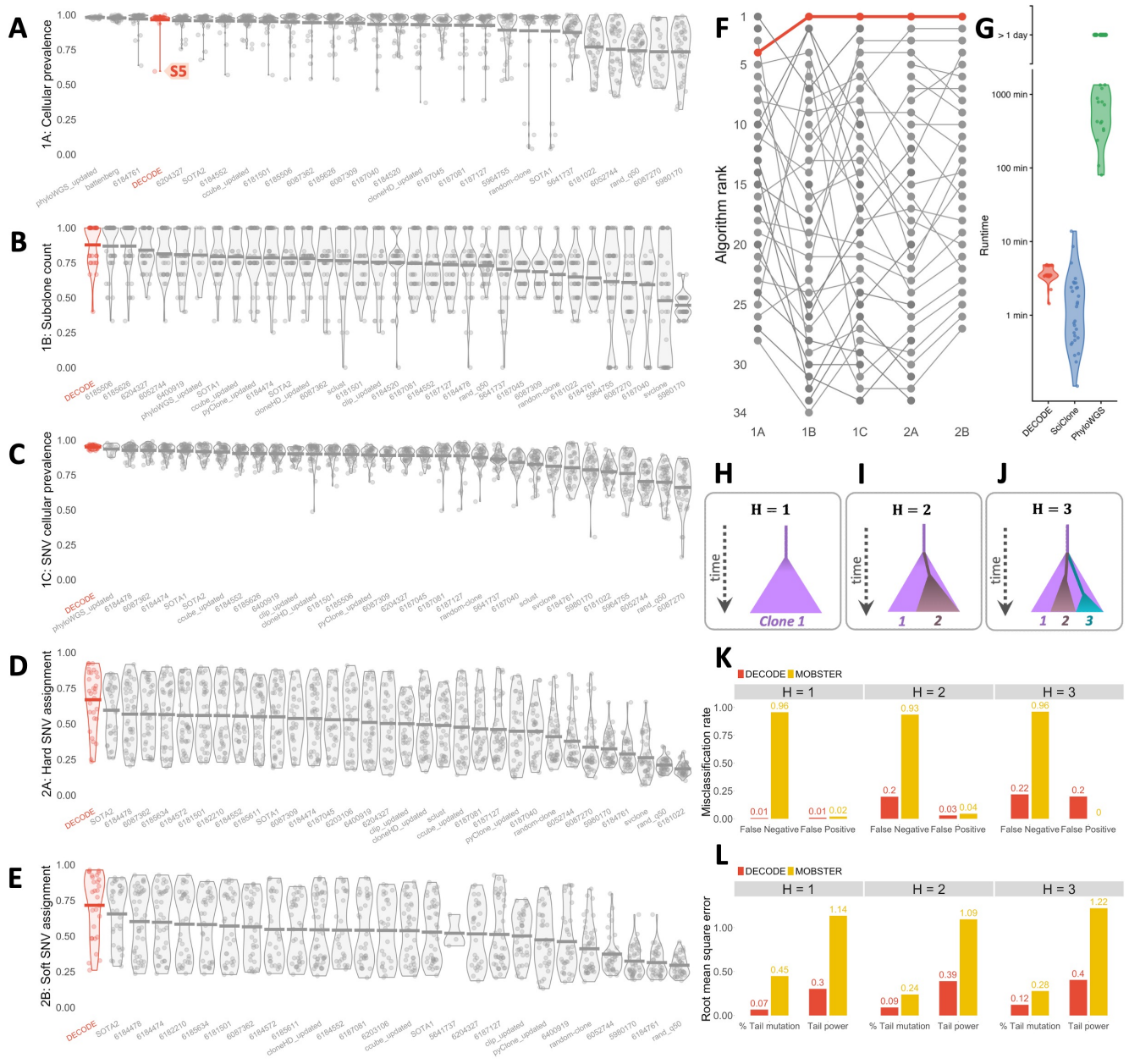


Figure 2: Synthetic benchmarking results. **A-G**: Results from benchmarking DECODE’s performance against 31 existing subclonal reconstruction methods analyzed in ICGC-TCGA DREAM Somatic Mutation Calling Tumor Heterogeneity and Evolution Challenge. **A-E**: DREAM score distributions for DECODE (red) and other algorithms (gray) in inferring cellular prevalence (**A**), subclone count (**B**), subclonal cellular prevalences and mutation counts (**C**), and deterministic and probabilistic mutation assignments (**D**, **E**). Each violin plot illustrates one algorithm’s scores across synthetic tumors (dots). Algorithms are ranked based on average scores across all samples (horizontal bars). **F**: Ranking of DECODE (red) and other algorithms (gray) across all DREAM tasks, based on average scores in each task. Each line connects available results from an algorithm. **G**: Comparison of computational runtime between DECODE, SciClone and PhyloWGS for analyzing DREAM samples. Each violin plot illustrates an algorithm’s runtime across synthetic tumors (dots). Runs that remained unfinished after 10^5 seconds are marked as incomplete (“>1 day”). **H-L**: Results from benchmarking DECODE’s tail detection and characterization against MOBSTER. **H-J**: Schematics of clonal evolution models for simulations with $H = 1$ (**H**), 2 (**I**) and 3 subclones (**J**). **K**, **L**: Misclassification rates (**K**) and root mean square errors (RMSEs) of estimated tail power and mutation proportion (**L**) from DECODE and MOBSTER across the three simulation batches **H-J**.

previous best results, respectively. This seems unexpected, given that DECODE only analyzes mutations with the same ploidy, and therefore may overlook subclones associated with copy number variations (CNVs). The results suggest that DECODE is sensitive enough to detect CNV-driven subclonal expansions, given that they harbor sufficient mutations in their initiating cells to form an observable cluster in the SFS.

Compared with existing algorithms, DECODE performs particularly well in assigning mutations to individual subclones in both deterministic and probabilistic configurations (Fig. 2D, E), with 18% and 17% improvements compared to previous best methods, respectively. Therefore, we expect DECODE to accurately identify potential driver events for each subclone in single samples. Moreover, applications of DECODE in large-scale studies have the potential to reliably reveal frequent clonal and subclonal driver mutations.

Samples T3 and P5 in the DREAM Challenge provide some insight into DECODE’s performance (Supplementary Figs. 2, 3). The ground truth SFS for T3 consists of a truncal cluster at VAF $p_1 = 0.3$ and a small cluster at $p_2 = 0.18$, the latter obscured by a large cluster at $p_3 = 0.1$ (Supplementary Fig. 2A). The substantial overlap among these clusters leads many existing methods to identify only the two dominant ones, which underestimates the heterogeneity in the sample (Supplementary Fig. 2B). In contrast, DECODE detects that the dispersions of these clusters, governed by the coverage distribution, are not wide enough to cover the intermediate VAF region. It thus correctly identifies three embedded clusters at frequencies $p_1 = 0.29$, $p_2 = 0.19$ and $p_3 = 0.1$ (Supplementary Fig. 2C). On the other hand, sample P5 consists of a truncal cluster at $p_1 = 0.47$, a subclonal cluster at $p_2 = 0.19$ and a peak formed of false positives in the low-VAF region (Supplementary Fig. 3A). SciClone [Miller et al., 2014] infers five subclones, two of which overfit the truncal cluster (Supplementary Fig. 3B). DECODE accurately detects two clusters at $p_1 = 0.46$ and $p_2 = 0.18$, in addition to a third group at $p_3 = 0.05$ consisting of the false positives (Supplementary Fig. 3C). Together, these two examples underscore DECODE’s enhanced accuracy in detecting overlapping clusters, reducing the risk of either underfitting or overfitting cancers’ heterogeneity.

Notably, unlike most methods which excel on only a few objectives, DECODE performs strongly across all DREAM tasks (Fig. 2F). This is likely a result of its incorporation of sample-specific coverage in the SFS formulation and rigorous framework to select the parsimonious SFS deconvolution, which together enhance DECODE’s cluster detection. Across the 29 DREAM samples analyzed, SciClone [Miller et al., 2014], a popular clustering software, finishes in two minutes on average per sample (Fig. 2G). PhyloWGS [Deshwar et al., 2015], currently the state of the art (Fig. 2A-E), exceeds 1 day for 13 samples and requires 100 minutes or more in the remaining 16 samples (Fig. 2G). PhyloWGS’s long runtime is likely a result of its Markov chain Monte Carlo implementation, which is inherently difficult to parallelize, and its simultaneous analysis of both mutations and copy number aberrations, the latter not being utilized by DECODE or SciClone. DECODE is comparable to SciClone at roughly four minutes on average per sample, and is therefore applicable for large-scale genomic studies.

We further benchmarked DECODE’s detection and characterization of the SFS neutral tail against MOBSTER, currently the only other method that incorporates this component to our knowledge [Caravagna et al., 2020]. We simulated data comprising 1 – 3 clones (Fig. 2H-J). For each clone count, we created 1,800 synthetic tumors with tails (“T-type”) and 200 simulated samples without tails (“NT-type”). Each sample was simulated with purity, mutation burden, and sequencing coverage drawn at random from The Cancer Genome Atlas (TCGA) [Weinstein et al., 2013]. DECODE detects the neutral tail more reliably than MOBSTER, with lower false negative rates (i.e., missed tails in T-type samples) and false positive rates (i.e., spurious tails in NT-type samples) across all clone counts, with the exception of the false positive rate in $H = 3$ (Fig. 2K). This likely reflects MOBSTER’s tendency to detect the tail confidently only at high coverage ($> 100\times$), and to be more conservative at the lower coverage typical of TCGA [Caravagna et al., 2020, Salcedo et al., 2025]. Furthermore, when restricted to each method’s true positives (i.e., T-type samples in which the tail is successfully detected), DECODE’s estimates

of the tail power and the fraction of mutations belonging to the tail are also significantly more accurate than MOBSTER’s (Fig. 2L). This highlights the need to account for sample-specific coverage and mutation-calling biases, which substantially distort the SFS tail.

We next investigated which attributes of genomic data most strongly affect the accuracy of DECODE’s inferred tail power and mutation proportion (Supplementary Figs. 4-6). As can be expected, DECODE’s results tend to be more accurate for samples with higher purity, coverage depth and mutation count. Nevertheless, only two properties consistently affect DECODE’s tail characterization at a significant level across all clone counts. The first is cluster count error: over- or under-estimating the clonality distorts the tail compartment, leading to inaccurate tail parametrization (Supplementary Figs. 4-6). However, this is unlikely to affect many clinical applications, as DREAM results indicate that DECODE’s inferred clonality for realistic samples is generally reliable (Fig. 2B). The second significant factor is entropy in the tail region (Supplementary Figs. 4-6). High entropy, caused by low-*VAF* subclones obscuring the tail, renders tail parameters non-identifiable. Hence, for samples with low purity, high clonality or rare subclones, we expect DECODE’s tail characterization to be less reliable.

Together, the results from the DREAM Challenge and our SFS tail benchmark study suggest that DECODE is well positioned to become a new standard in characterizing tumor clonality and evolution mode from individual genomic samples, with potential for clinical applications.

2.3 Incorporating the neutral tail enhances the reconstruction of clonal evolution in acute myeloid leukemia

While DECODE was designed to decompose single genomic samples, we also explored its utility for multi-sample data by analyzing paired acute myeloid leukemia (AML) whole-genome sequences at diagnosis (Dx) and relapse (Rx) from 10 patients in [Shlush et al., 2017]. Multi-sample deconvolution with SciClone [Miller et al., 2014] infers 4-7 subclones per patient (Fig. 3A, Supplementary Figs. 7-15). Analyzing each of the 20 samples individually, MOBSTER distinguishes 2-3 subclones per sample (Fig. 3B-C, Supplementary Figs. 7-15) but detects the neutral tail in only one sample, which was sequenced at $137\times$ (Supplementary Fig. 14B). MOBSTER’s inability to identify the tail in 19/20 cases is likely because the sequencing depths of these samples (median $54\times$, range $45 - 70\times$) are significantly lower than the $> 90 - 100\times$ threshold recommended by its authors. Meanwhile, DECODE identifies neutral tails in all 20 samples, in addition to 1 – 3 subclones per sample (Fig. 3D-E, Supplementary Figs. 7-15). We investigated the evolution models extrapolated from DECODE’s tail-aware fits, in comparison to SciClone’s joint clustering results.

In Patient 1, SciClone detects one subclone present only at diagnosis, another subclone specific to relapse, and three subclones shared at both time points (Fig. 3A). Intriguingly, the mean *VAFs* of the shared subclones remain unchanged from diagnosis to relapse (Fig. 3F). The implication of clonal selection equilibrium may be biologically questionable, given that the intervening treatment resulted in clinical and morphological remission in this patient [Shlush et al., 2017]. In contrast, DECODE determines that the mutations with $VAF \approx 0.5$ in both samples (i.e., SciClone’s cluster A) form the truncal cluster, while the other two shared subclones and both Dx- and Rx-specific subclones inferred from SciClone (i.e., clusters B, C, D and E) instead form the neutral tails in each sample (Fig. 3G).

DECODE’s truncal cluster mutation assignments are highly consistent across samples, despite each sample being decomposed independently (Fig. 3H). Likewise, the majority of shared mutations with lower *VAFs* are assigned to the tail in both samples, with the mutations specific to each time point making up the remainder of each tail. We therefore hypothesize that the diagnosis sample consists of cells proliferating at similar rates, and the

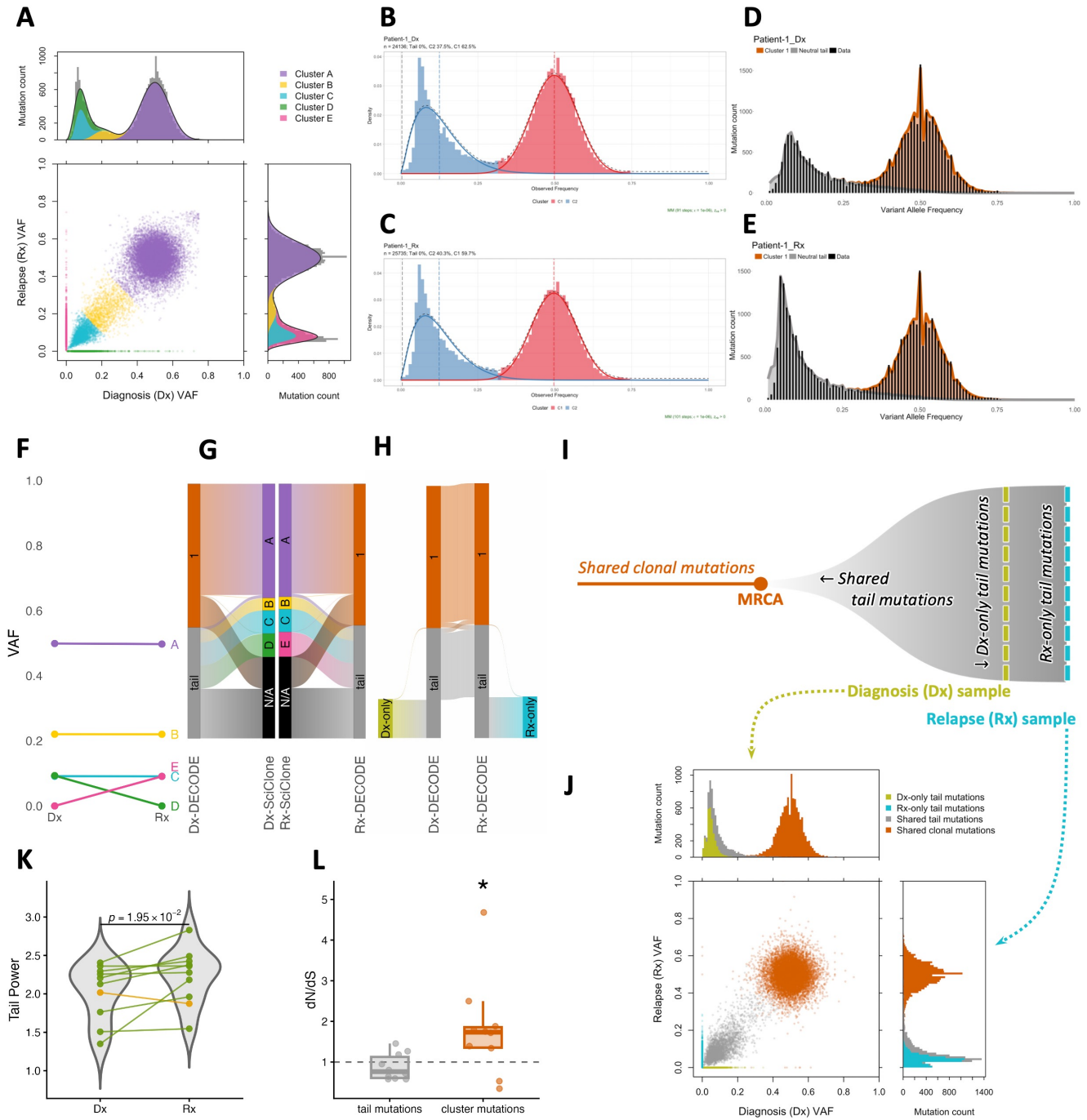


Figure 3: DECODE reveals neutral evolution in longitudinal acute myeloid leukemia. **A-J**: Mutation clustering for patient AML-1. **A**: SciClone's joint clustering of mutations at diagnosis (Dx) and relapse (Rx). **B-E**: Deconvolution of mutations using MOBSTER (**B**, **C**) and DECODE (**D**, **E**) for each sample at diagnosis (**B**, **D**) and relapse (**C**, **E**). **F**: Mean VAFs between diagnosis and relapse of subclones inferred from SciClone. **G**: Alluvial diagrams of subclonal assignments for mutations at diagnosis (left) and relapse (right) between DECODE and SciClone. N/A: mutations without SciClone assignments. Widths are proportional to mutation fractions. **H**: Alluvial diagram of DECODE subclonal assignments for mutations at diagnosis and relapse. Outer bars contain mutations unique to each sample. **I**: Proposed model of evolution for patient AML-1. **J**: Synthetic multi-sample data simulated with TEMULATOR under the evolutionary model in **I**. **K**: Comparison of tail powers α at diagnosis and relapse inferred from DECODE. Lines connecting α 's for the same patients are highlighted in green if α is higher in Rx, and yellow otherwise. p -value from one-sided Wilcoxon test. **L**: dN/dS analysis of mutations assigned to tails and clusters in each patient (dots). *: p -value from one-sided Wilcoxon test < 0.05 (alternative: dN/dS > 1).

relapse sample simply reflects a continuation of their expansion following a treatment-induced bottleneck (Fig. 3I). Synthetic multi-sample data, simulated under this evolution model with TEMULATOR [Heide, 2022] (Fig. 3J) are indeed consistent with the joint SFS pattern observed in Patient 1’s data (Fig. 3A). In this evolution model, the mutations present in the primary tumor’s MRCA form a shared clonal cluster, centered at VAFs determined by the cellular prevalence at each time point. Mutations arising soon after the MRCA exist in relatively high cell fractions, thus they comprise a portion of the diagnosis SFS tail that persists at relapse. The shared tail mutations therefore form a cloud in the joint SFS, distributed along the diagonal from the low-VAF corner to the shared clonal cluster. In contrast, late-arising mutations exist in low cell fractions at diagnosis and are prone to imminent extinction, contributing a Dx-specific tail at lower VAFs than the shared tail. Finally, mutations emerging after diagnosis constitute the Rx-specific SFS neutral tail. This evolution model is consistent with the original publication, which found that the diagnosis and relapse blasts in this patient are highly related [Shlush et al., 2017]. The joint SFS of 6/10 patients mirror Patient 1’s; in each case, DECODE’s deconvolutions consist only of neutral tails and truncal clusters (Supplementary Figs. 8, 11, 12, 13, 15). We therefore hypothesize that the evolution of these tumors follows the model in Fig. 3I. DECODE detects subclones in the remaining 4 cases, three of which exhibit higher clonality at relapse (Supplementary Figs. 7, 9, 10).

Previous work in population genetics has revealed that the SFS tail power α is an indicator of ongoing tumor expansion, with $\alpha = 1$ in cell populations of constant size [Watterson, 1975] and $\alpha = 2$ in exponentially expanding populations [Griffiths and Tavaré, 1998]. DECODE results demonstrate an increase in tail power at relapse in 9/10 AML patients (Fig. 3K), suggesting that cancer cells expand more aggressively at relapse than at diagnosis, even in patients with no detectable subclones. Finally, dN/dS analysis confirms that mutations that DECODE assigns to the SFS tail are consistent with neutral evolution (i.e., dN/dS ≈ 1) while the variants assigned to mutation clusters indicate positive selection (i.e., dN/dS > 1), further corroborating DECODE’s detection of neutral tails and clonality in the SFS (Fig. 3L).

This study underscores three advantages of applying DECODE to analyze cancer genomic data. First, it is able to identify neutral tails in the SFS even for samples with low sequencing depths, thanks to our correction for coverage distribution and mutation filtering. Second, by accounting for mutations in the SFS tail, it avoids inferring spurious subclones from neutrally accumulating passenger mutations. This often results in simpler and more biologically plausible reconstructions of tumor evolution. Finally, DECODE’s inferred tail powers provide an orthogonal measure for characterizing single cancer samples, independent from clonality. SFS tail power reflects the ongoing tumor expansion mode, which has been observed to be a strong biomarker for patient survival in time-series experiments [Stein et al., 2008]. By enabling the estimation of tumor expansion rates from single genomic samples, DECODE has the potential to support clinical decision-making in settings where serial or multi-region sampling is impractical.

2.4 DECODE infers clonality, tail power and sample purity accurately and consistently for pan-cancer samples

We implemented DECODE to analyze pan-cancer whole-genome sequencing (WGS) samples from The Cancer Genome Atlas (TCGA) [Weinstein et al., 2013]. We first examined the consistency of DECODE’s results across same-patient biological replicates, sampled from different tumor regions across multiple cancer types. DECODE’s assignment of mutations to the tail or specific SFS clusters is more concordant across 57 multi-region WGS pairs, compared to SciClone’s subclonal assignments (Fig. 4A; median adjusted Rand index (ARI): 0.19 in DECODE, 0.09 in SciClone). Because samples from the same patient are expected to share a broadly common evolutionary

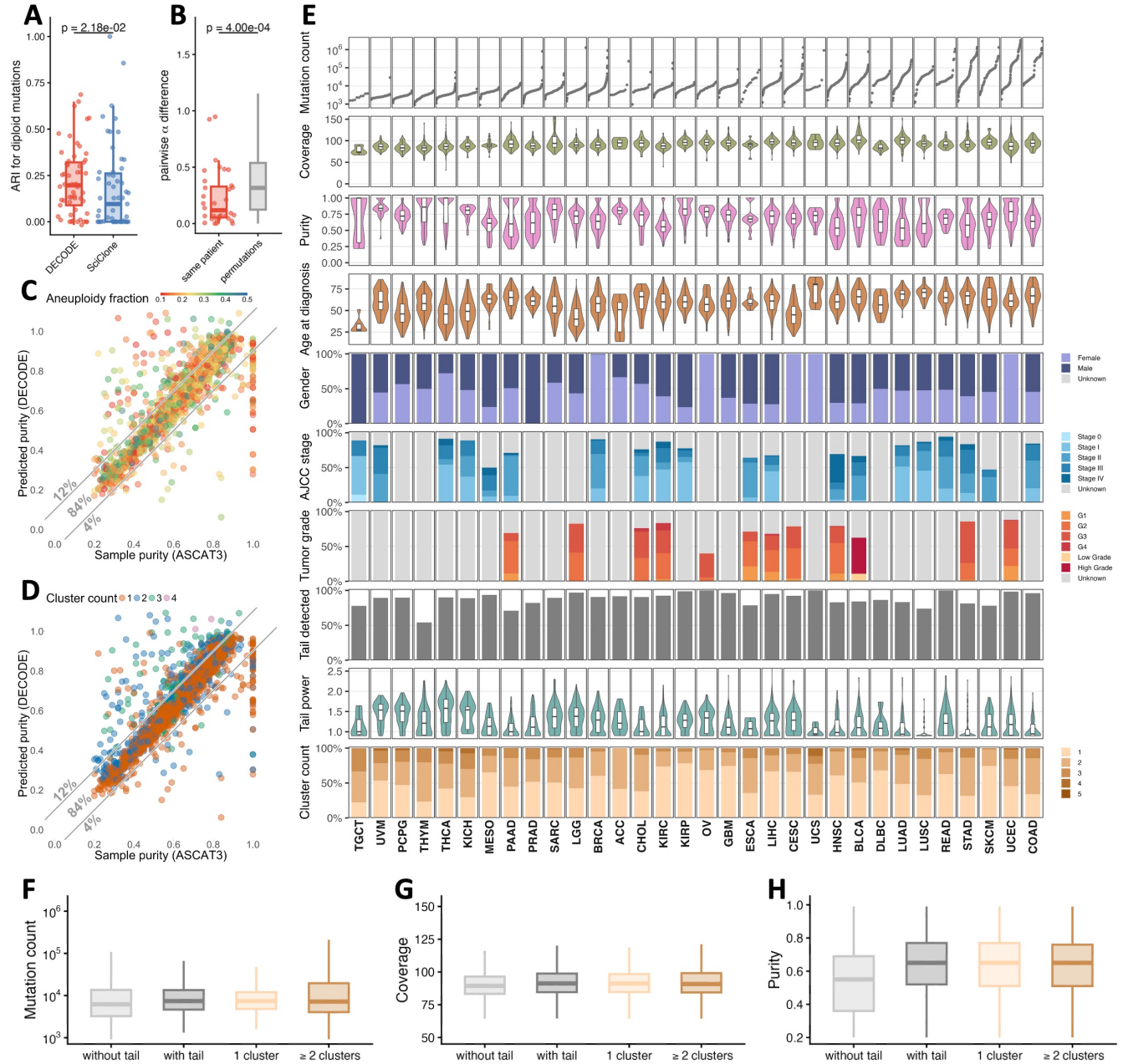


Figure 4: Deconvolutions of pan-cancer samples in TCGA. **A**: Comparison of ARI scores for deconvolutions of same-patient biological replicates in TCGA between DECODE and SciClone. p -value from one-sided Wilcoxon test. **B**: Comparison of differences in DECODE-inferred tail powers among same-patient biological replicates and permuted samples. p -value from permutation test. **C**, **D**: Comparison of sample purities estimated by ASCAT3 against those predicted from DECODE-inferred truncal clusters, with colors indicating sample aneuploidy fraction (**C**) or DECODE-inferred cluster count (**D**). ASCAT3 assigns purity = 1 for samples where aneuploidy fractions are too low for accurate estimation. Diagonal lines demarcate correctly predicted purities (error < 0.1) from over- and under-estimates; percentages indicate the proportion of samples in each group. **E**: Summary of tumor characteristics and DECODE results across TCGA cancer types, ordered by median mutation burden. Top to bottom: sample-specific mutation burdens; whole-genome sequencing depths; ASCAT3-inferred sample purity; patient ages at diagnosis; patient genders; tumor stages per the American Joint Committee on Cancer (AJCC) system; tumor grades; proportions of samples in which DECODE detected neutral tails; DECODE-inferred tail powers; and DECODE-inferred cluster counts. **F-H**: Distributions of mutation burden (**F**), sequencing depth (**G**) and sample purity (**H**) across all samples, stratified by whether DECODE detected a tail and its inferred cluster count.

history, the concordance indicates that DECODE recovers reproducible clonal structure rather than sample-specific noise.

Additionally, among 35 pairs of biological replicates where DECODE detects a tail in both samples, the inferred tail powers agree significantly more closely within the same patients than across permuted pairs (Fig. 4B). Given that the AML study establishes the importance of the tail power as an indicator of tumor severity (Fig. 3K), this agreement further reaffirms the robustness of the tail power as a measure of tumor expansion.

As in the DREAM Challenge (Fig. 2A), cellular prevalence can be predicted from the truncal cluster VAF. In 84% of the TCGA samples with aneuploidy fraction > 0.1 ($n = 1,966/2,333$), the predictions are within 10% of the sample purities inferred from ASCAT3 [Van Loo et al., 2010] (Fig. 4C, D). Among samples where DECODE’s prediction error exceeds 10%, roughly one in four are underestimates. This mainly results from the low level of aneuploidy in these samples, due to which ASCAT3 and other copy number-based methods cannot estimate purity accurately (Fig. 4C). DECODE overestimates the cellular prevalence in 12% of the samples, indicating that alternative fits with fewer clusters and smaller truncal cluster VAFs would have been more appropriate for these samples (Fig. 4D). We note that errors in allele-specific copy number calling can introduce spurious high-VAF clusters in the SFS. One such case is copy-neutral loss of heterozygosity miscalled as heterozygous: in a sample of purity ρ , truncal mutations on the retained allele then appear at VAF ρ rather than the expected $\rho/2$, mimicking a subclone at high cellular prevalence and rendering the maximum cluster VAF an unreliable predictor for purity. Overall, copy number-based methods remain the most robust approach for estimating sample purity from genomic data. However, for samples with low aneuploidy where these methods have limited signal (e.g., 45% samples in TCGA have aneuploidy fraction ≤ 0.1), the strong agreement with ASCAT3 supports DECODE as a complementary route to estimating cellular prevalence.

We evaluated DECODE’s deconvolutions of primary tumors in TCGA ($n = 4,058$ patients across 32 cancer types). DECODE detects the neutral tail in 89% of the samples, with cancer-specific proportions ranging from 54% in thymomas (THYM) to 100% of ovarian cancers (OV), rectum adenocarcinomas (READ) and uterine carcinosarcomas (UCS) (Fig. 4E). In addition, it infers the existence of multiple mutation clusters in 47% of the patients ($n = 1,898/4,058$), with the proportion of samples with two or more clusters ranging from 22% in kidney renal papillary cell carcinomas (KIRP) to 78% in testicular germ cell tumors (TGCT). The inferred tail powers fall between 1 and 2 for the majority of patients (67%, $n = 2,433/3,625$), with only 1% of samples ($n = 51/3,625$) having tail powers higher than 2. As noted above, this indicates that most tumors occupy a continuum between stable size ($\alpha \approx 1$) and exponential expansion ($\alpha \approx 2$).

DECODE’s inference of clonality does not seem to be affected by mutation burden, sequencing coverage or sample purity (Fig. 4F-H). Its detection of neutral tail in the SFS likewise does not depend on variant count or read depth (Fig. 4F-G). However, samples where DECODE cannot identify a tail have significantly lower purity than those where it does (Fig. 4H). This is because, at low cellular prevalence, the mutation clusters are compressed into the low-VAF region in the SFS where the tail resides. The lack of a clear boundary between tail and clusters then leads DECODE to prefer a more parsimonious fit without a tail.

2.5 DECODE clonality stratifies overall survival in low-grade glioma

We assessed whether DECODE-inferred clonality is associated with patient outcome. We defined TCGA cohorts by cancer type, and further stratified each cancer type by histological stage or tumor grade. Stratifying patients in each cohort into two subgroups by the median DECODE-inferred cluster count reveals significant survival differences in two cohorts (Fig. 5A). The first is the primary low-grade glioma (LGG) cohort, where patients with

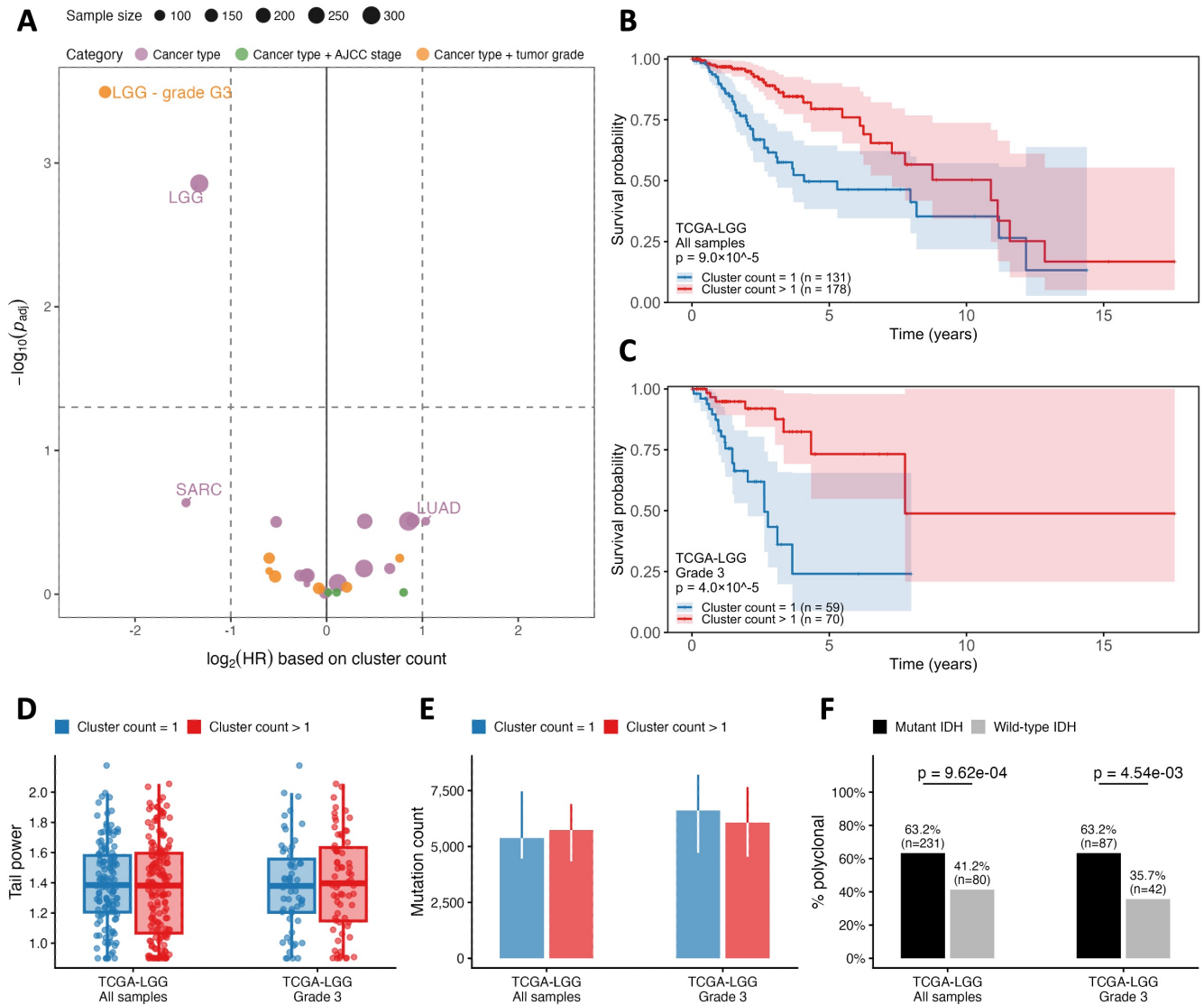


Figure 5: Prognostic value of DECODE-inferred clonality in TCGA. **A**: Volcano plot summarizing the association between cluster count and overall survival across TCGA cohorts, defined by cancer type alone, cancer type with AJCC pathological stage, or cancer type with tumor grade. Samples within each cohort were split at the median cluster count; the x-axis shows the hazard ratio (HR) between the resulting subgroups, and the y-axis shows the Benjamini-Hochberg adjusted log-rank p-value p_{adj} comparing their Kaplan-Meier curves. Dashed lines mark $p_{adj} = 0.05$, and $HR = 2$ and $1/2$. **B**, **C**: Kaplan-Meier overall survival curves for the TCGA lower-grade glioma (LGG) cohort, comprising all samples (**B**) or only grade 3 tumors (**C**), partitioned at the median cluster count ($= 1$ in both). p-values from log-rank tests comparing subgroups in each cohort. **D**, **E**: DECODE-inferred tail powers (**D**) and mutation burdens (**E**) in the LGG and LGG grade 3 cohorts, partitioned at the median cluster count ($= 1$). **F**: Proportions of samples with cluster count > 1 in the LGG and LGG grade 3 cohorts, partitioned by IDH status. p-values from two-sided Fisher's exact test.

polyclonal tumors (two or more subclones) survive significantly longer than those with monoclonal tumors (Fig. 5B; median survival time: 4.1 years in samples with one cluster, 10.9 years in samples with two or more clusters). Similarly, LGG grade 3 patients with multiple detected subclones survive on average roughly three times longer compared to those with clonally homogeneous tumors (Fig. 5C; median survival time: 2.6 years in samples with

one cluster, 7.8 years in samples with two or more clusters). The fact that cluster count is clinically meaningful within the LGG grade 3 cohort indicates that for some cancer types, DECODE-inferred clonality may provide prognostic information beyond what histology and staging capture.

We investigated the biological factors driving the survival differences uncovered by DECODE clonality in LGG. There is no significant difference in tail power or mutation count between patient subgroups in either the LGG or LGG grade 3 cohort (Fig. 5D-E), suggesting that the survival disparity is unlikely to reflect differences in tumor expansion mode or mutation burden. However, DECODE detected multiple subclones significantly more often in IDH-mutant than in IDH-wildtype tumors in both cohorts (Fig. 5F). IDH-mutant LGGs are known to grow more slowly than IDH-wildtype tumors [van den Bent et al., 2024] and to confer substantially longer overall survival [The Cancer Genome Atlas Research Network, 2015]. We therefore hypothesize that IDH-mutant LGG’s longer evolutionary timescale gives subclones more time to diversify and reach detectable VAF, potentially under near-neutral dynamics [Williams et al., 2016], which would account for the strong association between higher ITH and favorable outcome (Fig. 5B-C). Although the association between IDH mutation and elevated ITH has not been reported in LGG to our knowledge, mutant IDH has been linked to increased clonality in intrahepatic cholangiocarcinoma [Xiang et al., 2021] and acute myeloid leukemia [Sirenko et al., 2025].

3 Discussion

DECODE is a novel mutation clustering method that builds on our previous work in population genetics and parameter inference [Dinh et al., 2020, 2025]. The synthetic tests showed that DECODE can detect and characterize the SFS neutral tail accurately. Furthermore, DECODE estimates multiple aspects of tumor clonality more reliably and consistently than current approaches (Fig. 2). Incorporating the tail in the SFS simplifies the decomposition of many acute myeloid leukemias, yielding a more biologically realistic model of how these cancers evolve toward relapse (Fig. 3). Applied to TCGA pan-cancer data, DECODE reveals an SFS tail in the majority of samples across tumor types (Fig. 4), and its clonality estimates reveal a correlation between ITH and survival in low-grade glioma (Fig. 5). Overall, DECODE’s clonality and tail-power inferences provide two complementary axes for estimating tumor clonality and expansion mode from genomic data, both of which have been shown to be important biomarkers for patient outcome [Swanson et al., 2008, Ferté et al., 2014, Andor et al., 2016, Caswell and Swanton, 2017, Dagogo-Jack and Shaw, 2018].

Despite these advantages, DECODE has several areas for improvement. First, because it relies on the VAF-based SFS, DECODE’s clonality inference is valid only for variants sharing the same copy-number state. It is therefore suitable to analyze most diploid tumors, but not those with increased chromosomal instability. Second, DECODE was designed to analyze individual samples. Unlike PhyloWGS and SciClone [Deshwar et al., 2015, Miller et al., 2014], among others, it currently cannot jointly analyze multi-region or longitudinal samples from the same patient. However, our AML study showed that DECODE’s single-sample results can be merged to reconstruct the overall clonal evolution across multiple samples (Fig. 3). Finally, DECODE does not explicitly infer the phylogenetic relationships between subclones. Applying constraints such as the pigeonhole principle, integrating tree-building approaches [Deshwar et al., 2015], or extending DECODE to multi-sample data may enable reconstruction of the subclonal phylogeny.

Methods

A DECODE’s framework

A.1 Expected site frequency spectrum given coverage distribution and mutation filtering

We assume that a tumor sample consists of H subclones, and that the fraction c_h of sampled cells belonging to subclone h share the ω_h mutations present before the subclone’s most recent common ancestor (MRCA) (Fig. 1A). We denote the ancestral clone with $h = 1$ from which other subclones $h = 2, \dots, H$ originate, so that c_1 constitutes the cellular prevalence (i.e., sample purity). Assuming diploidy, the expected number of mutations present in k copies among n alleles from a sample of $n/2$ cells, following results in population genetics [Griffiths and Tavaré, 1998, Durrett, 2013, Dinh et al., 2020], is:

$$\mathbb{E}(S^n(k)|\theta_H) \approx \omega_0 \cdot g_{\text{tail}}(k|\alpha) + \sum_{h=1}^H \omega_h \cdot g_{\text{peak}}(k|p_h) \quad (2)$$

where $\theta_H = \{\alpha, p_1, \dots, p_H, \omega_0, \dots, \omega_H\}$ are the parameters to be inferred. Mutations in cluster h follow a binomial distribution with mean variant allele frequency (VAF) $p_h = c_h/2$ [Tarabichi et al., 2021]:

$$g_{\text{peak}}(k|p_h) = \binom{n}{k} p_h^k (1 - p_h)^{n-k}$$

The ω_0 mutations in the “foreground” occur neutrally in all subclones, following a tail distribution with power α :

$$g_{\text{tail}}(k|\alpha) \sim \frac{1}{k^\alpha}$$

for $k = 1, \dots, n \cdot p_1$. Previous works have shown that for well-mixed samples, the neutral mutation tail is of power $\alpha = 1$ in a population of constant size [Watterson, 1975], and $\alpha = 2$ in an exponentially expanding population [Griffiths and Tavaré, 1998]. Different modes of tumor growth, spatial effects and other confounding factors can have an impact on α .

We previously developed the formula for the expected observed SFS that accounts for the sequencing coverage and data filtering [Dinh et al., 2020], relevant details of which are described here. We define the coverage as the probability mass function φ , such that $\varphi(r)$ is the probability that a site is covered by a total of r reads. Because DNA-seq samples typically contain $\geq 10^5$ cells and are sequenced with mean coverage $< 100\times$, we assume that the r reads originate from different cells. For a mutation present in k alleles in the sample, its exact VAF is k/n , therefore the number of variant reads follows $z|r \sim \text{Binomial}(r, k/n)$. The expected mutation count with observed frequency $f \in (f_1, f_2]$, for $0 \leq f_1 < f_2 \leq 1$, is then:

$$\mathbb{E}(S((f_1, f_2])|\theta_H) = \sum_{k=1}^n \mathbb{E}(S^n(k)|\theta_H) \sum_r \varphi(r) \cdot \mathbb{P}\left(\text{Binomial}\left(r, \frac{k}{n}\right) \in (f_1 r, f_2 r]\right)$$

where $\mathbb{P}(\text{Binomial}(r, \frac{k}{n}) \in (f_1 r, f_2 r])$ is the probability that a mutation covered by r reads with true VAF k/n has observed VAF $f = z/r \in (f_1, f_2]$.

The sequencing data can be represented as $\mathcal{D} = \{(z_i, r_i)\}_{i=1,2,\dots}$, where the i th mutation is observed in z_i variant reads among r_i total reads. With thresholds $L, M \geq 0$ for variant and total read counts, respectively, the

filtered data is $\mathcal{D}^{[L,M]} = \{(z_i, r_i) \in \mathcal{D} : z_i \geq L, r_i \geq M\}$. The observed SFS $\bar{S}^{[L,M]}$ can then be tabulated from $\{z_i/r_i : (z_i, r_i) \in \mathcal{D}^{[L,M]}\}$ with bin size b . The expected observed SFS, accounting for the mutation filtering, is:

$$\mathbb{E} \left(S^{[L,M]}((f_1, f_2)) \middle| \theta_H \right) = \sum_{k=1}^n \mathbb{E} (S^n(k) | \theta_H) \sum_r \varphi^{[L,M]}(r) \cdot \mathbb{P} \left(\text{Binomial} \left(r, \frac{k}{n} \right) \in (f_1 r, f_2 r] \cap [L, \infty) \right) \quad (3)$$

for each bin $(f_1, f_2] = (0, b], (b, 2b], \dots$. The coverage distribution can be approximated directly from normalized counts of mutations with given total read counts in the filtered mutation data:

$$\varphi^{[L,M]}(r) = \begin{cases} \frac{|\{(z_i, r_i) \in \mathcal{D}^{[L,M]} : r_i = r\}|}{|\mathcal{D}^{[L,M]}|} & \text{if } r \geq M, \\ 0 & \text{if } r < M \end{cases}$$

where $|\cdot|$ counts the mutations in a dataset. Together, Eqs. (2) and (3) form a connection between the exact but unobserved SFS, which contains signals for the neutral tail and mutation clusters, and the expected observed SFS, which accounts for sample-specific sequencing coverage and mutation filtering. By default, DECODE assumes a sample with $n = 1000$ alleles and the SFS is computed with $N_{\text{bins}} = 100$ bins of size $b = 0.01$.

A.2 Inferring parameters for the neutral tail and mutation clusters

Assuming that the cluster count H is known, DECODE seeks $\theta_H = \{\alpha, p_1, \dots, p_H, \omega_0, \dots, \omega_H\}$ that best fits the observed mutational data. To ensure reliability, it analyzes data with two distinct filtering strategies $(L^{[A]}, M^{[A]})$ and $(L^{[B]}, M^{[B]})$ simultaneously. By default, DECODE selects $L^{[A]} = L^{\min} + \max(1, \frac{1}{4}(L^{\max} - L^{\min}))$ and $L^{[B]} = L^{\min} + 3 \cdot \max(1, \frac{1}{4}(L^{\max} - L^{\min}))$, where L^{\min} and L^{\max} are the minimum and 95th percentile of the variant read counts observed in the data, respectively. These thresholds filter out the mutations with the fewest variant reads, as they are most affected by mutation calling biases. Furthermore, increasing L changes the shape of the SFS, especially the neutral tail which mainly exists in the low-VAF region. The total read count thresholds $M^{[A]}$ and $M^{[B]}$ are then chosen such that the correspondingly filtered data subsets $\mathcal{D}^{[A]}$ and $\mathcal{D}^{[B]}$ contain roughly the same numbers of mutations, to avoid overfitting either subsample. DECODE is therefore required to fit the distinct SFS resulting from both $(L^{[A]}, M^{[A]})$ and $(L^{[B]}, M^{[B]})$ well, hence increasing the robustness of its results.

Following Eqs. (2)-(3), the expected number of tail mutations observed in $\mathcal{D}^{[A]}$ is $\omega_0^{[A]} = \gamma_0^{[A]} \cdot \omega_0$, where

$$\gamma_0^{[A]} = \sum_{k=1}^n g_{\text{tail}}(k|\alpha) \sum_r \varphi^{[L^{[A]}, M^{[A]}]}(r) \cdot \mathbb{P} \left(\text{Binomial} \left(r, \frac{k}{n} \right) \in [L, \infty) \right)$$

is a constant. Similarly, the expected number of mutations in cluster h in $\mathcal{D}^{[A]}$ is $\omega_h^{[A]} = \gamma_h^{[A]} \cdot \omega_h$, where

$$\gamma_h^{[A]} = \sum_{k=1}^n g_{\text{peak}}(k|p_h) \sum_r \varphi^{[L^{[A]}, M^{[A]}]}(r) \cdot \mathbb{P} \left(\text{Binomial} \left(r, \frac{k}{n} \right) \in [L, \infty) \right)$$

Therefore, inferring θ_H is equivalent to estimating $\{\alpha, p_1, \dots, p_H, \omega_0^{[A]}, \dots, \omega_H^{[A]}\}$, which is more practical because $\omega_0^{[A]}, \dots, \omega_H^{[A]}$ are naturally limited by the filtered mutation count $|\mathcal{D}^{[A]}|$, providing unbiased bounds for parameter estimation. For a given parameter set, the true mutation counts can be back-calculated as $\omega_h = \omega_h^{[A]} / \gamma_h^{[A]}$.

To estimate $\theta_H = \{\alpha, p_1, \dots, p_H, \omega_0^{[A]}, \dots, \omega_H^{[A]}\}$, we implement Approximate Bayesian Computation sequential Monte Carlo via distributional random forests (ABC-SMC-DRF), a parameter inference method that we recently introduced which integrates distributional random forests (DRFs) in the framework of Approximate Bayesian Computation sequential Monte Carlo (ABC-SMC) [Dinh et al., 2025]. On the one hand, by implementing DRF, ABC-SMC-DRF circumvents the selection of statistical distance functions and tolerance thresholds in

traditional ABC methods, which may affect the result’s accuracy. On the other hand, by being embedded in the ABC-SMC framework, it focuses on the most likely regions in the parameter space iteratively, thus converging to the posterior distribution faster with less computation.

DECODE implements a modified version of ABC-SMC-DRF to accommodate cluster reordering. In the first iteration, 1000 samples of θ_H are drawn from the prior distribution:

$$\begin{aligned}\alpha &\sim \text{Uniform}(0.9, 5) \\ p_h &\sim \text{Uniform}(0.01, 1), \text{ for } h = 1, \dots, H \\ \omega_h^{[A]} &\sim \text{Uniform}\left(0, 2 \left| \mathcal{D}^{[A]} \right| \right), \text{ for } h = 1, \dots, H\end{aligned}$$

For each θ_H , the true mutation counts $\omega_0, \dots, \omega_H$ are computed, then the observed SFS is predicted from Eq. (3) for thresholds $(L^{[A]}, M^{[A]})$ and $(L^{[B]}, M^{[B]})$. The relative error for each SFS prediction $* = A, B$ compared to the correspondingly filtered SFS is defined as

$$E^{[*]} = \frac{1}{|\mathcal{D}^{[*]}|} \sum_{b=1}^{N_{\text{bins}}} \text{error}(b) \quad (4)$$

where

$$\text{error}(b) = \left| \bar{S}^{[L^{[*]}, M^{[*]}]}((b \cdot (n-1), b \cdot n)) - \mathbb{E}\left(S^{[L^{[*]}, M^{[*]}]}((b \cdot (n-1), b \cdot n)) \middle| \theta_H\right) \right| \quad (5)$$

is the error in bin b . The total error for parameter set θ_H is then $E = E^{[A]} + E^{[B]}$. DECODE trains a DRF on the 1000 particles of θ_H and corresponding errors, then derives weights for the particles to minimize the error [Ćevič et al., 2022]. Finally, it reorders the pairs $\{p_1, \omega_1^{[A]}\}, \dots, \{p_H, \omega_H^{[A]}\}$ in each θ_H such that $p_1 > \dots > p_H$.

In the second iteration, DECODE samples 1000 new θ_H ’s from the weighted particles in iteration 1. It then perturbs each parameter set with Gaussian kernels, where each kernel’s variance is twice the corresponding parameter’s empirical variance in iteration 1. This adaptive perturbation scheme has been shown to be unbiased and efficient in the context of ABC-SMC [Beaumont et al., 2009]. The new particles’ weights are then estimated with DRF and the clusters in each particle are sorted, similarly to iteration 1. DECODE continues until iteration 10, and the final weighted particle set approximates the posterior distribution for $\{\alpha, p_1, \dots, p_H, \omega_0^{[A]}, \dots, \omega_H^{[A]}\}$.

To ensure stability and accuracy, DECODE also performs a guided parameter inference if the results for θ_{H-1} are available. It guesses where cluster H resides by finding the bin b with central VAF $\hat{p}_H = b \cdot (n - \frac{1}{2})$ where the average error from Eq. (5) for $H - 1$ clusters is maximal, and posits that the mutation count for cluster H in $\mathcal{D}^{[A]}$ is approximately $\hat{\omega}_H = \left| \mathcal{D}^{[A]} \right| - \text{mean}_{\theta_{H-1}} \left[\mathbb{E}\left(S^{[L^{[A]}, M^{[A]}]}((0, 1)) \middle| \theta_{H-1}\right) \right]$, the average residual in the data compared to the SFS predicted with the previously inferred θ_{H-1} particles. The prior distribution for this inference is then

$$\begin{aligned}\alpha &\sim \mathcal{N}_{\text{tr}}(\hat{\alpha}, 0.05; 0.9, 5) \\ p_h &\sim \mathcal{N}_{\text{tr}}(\hat{p}_h, 0.05; 0.01, 1), \text{ for } h = 1, \dots, H \\ \omega_h^{[A]} &\sim \mathcal{N}_{\text{tr}}\left(\hat{\omega}_h^{[A]}, 0.05 \left| \mathcal{D}^{[A]} \right|; 0, 2 \left| \mathcal{D}^{[A]} \right| \right), \text{ for } h = 0, \dots, H\end{aligned}$$

where $\hat{\alpha}, \hat{p}_1, \dots, \hat{p}_{H-1}, \hat{\omega}_0^{[A]}, \dots, \hat{\omega}_{H-1}^{[A]}$ are the average estimates from θ_{H-1} , and $\mathcal{N}_{\text{tr}}(\mu, \sigma; a, b)$ is the normal distribution with mean μ and variance σ^2 , truncated to within (a, b) . DECODE then proceeds to estimate θ_H with the modified ABC-SMC-DRF procedure, similar to the inference from uniform prior distributions. Finally, the posterior distributions from uninformed and informed inferences are compared, and the result with lower average error is selected as DECODE’s final estimate for θ_H .

A.3 Parsimonious model selection

DECODE seeks the smallest cluster count H such that θ_H captures the observed mutational data. To ensure robustness, DECODE examines the SFS predicted by the inferred θ_H for a third filtering strategy $(L^{[V]}, M^{[V]})$, where $L^{[V]} = L^{\min} + 2 \cdot \max(1, \frac{1}{4}(L^{\max} - L^{\min}))$ and $M^{[V]}$ is the minimum total read count observed in the data. As $L^{[A]} < L^{[V]} < L^{[B]}$, DECODE can validate whether the θ_H distribution inferred for filters $(L^{[A]}, M^{[A]})$ and $(L^{[B]}, M^{[B]})$ can predict the novel SFS shape extracted from $(L^{[V]}, M^{[V]})$. For each θ_H particle, we define the Generalized Information Criterion [Konishi and Kitagawa, 1996]:

$$\text{GIC}(\theta_H) = E^{[V]} + \epsilon \cdot k_H \cdot \log |\mathcal{D}^{[V]}|$$

where $E^{[V]}$ is the relative error of the SFS prediction from Eq. (4), and $\epsilon = 0.0015$ calibrates the penalty of parameter count $k_H = 2H + 2\epsilon_{\text{tail}}$, relative to the mutation count $|\mathcal{D}^{[V]}|$ in the validation SFS subset, and $\epsilon_{\text{tail}} \in [0, 1]$ is the user-defined weight for neutral tail parameters.

After successive inferences with $H - 1$ and H clusters, DECODE quantifies whether the more complex model is justified by the improvement in prediction quality with a one-sided Wilcoxon rank sum test, where the null hypothesis is that $\text{median}\{\text{GIC}(\theta_{H-1})\} \leq \text{median}\{\text{GIC}(\theta_H)\}$. DECODE selects the model with H clusters and continues to infer θ_{H+1} if the p-value < 0.01 , otherwise it determines that the model with $H - 1$ clusters is the best fit for the data.

If there are rare subclones in the sample, the low-VAF clusters may obscure the tail, leading to unreliable SFS tail characteristics. Therefore, similar to MOBSTER [Caravagna et al., 2020], DECODE also fits the data with only mutation clusters, where the parameter set consists of $\theta'_H = \{p_1, \dots, p_H, \omega_1^{[A]}, \dots, \omega_H^{[A]}\}$. The inference of θ'_H distributions and selection of the parsimonious model are similar to DECODE routines with tail. DECODE chooses the overall best model between the selected results with H clusters and a tail, and H' clusters with no tail, with a one-sided Wilcoxon rank sum test for whether $\text{GIC}(\theta_H) \leq \text{GIC}(\theta'_{H'})$. The model with H' clusters and no tail is selected if the p-value < 0.01 , otherwise DECODE chooses the model with H clusters and a tail. In practice, we find $\epsilon_{\text{tail}} = 0.3$ to provide a good balance between ascertaining the tail parameters and selecting the clusters-only fit when the tail is too obscured.

B Synthetic testing

B.1 ICGC-TCGA DREAM Challenge

We tested DECODE’s performance using the synthetic benchmark from the ICGC–TCGA (International Cancer Genome Consortium – The Cancer Genome Atlas) DREAM Somatic Mutation Calling Tumor Heterogeneity and Evolution Challenge [Salcedo et al., 2020]. The dataset consists of 51 synthetic tumors, created to mimic real cancer samples under various assumptions of clonality, copy number aberrations, cellularity and sequencing depth, while also accounting for the tumors’ spatial growth. The authors simulated tumor BAM files with BAMSurgeon [Ewing et al., 2015], then identified somatic single-nucleotide variants (SNVs) and copy number aberrations (CNAs) with Genome Analysis Toolkit (GATK) MuTect [Cibulskis et al., 2013] and Battenberg [Nik-Zainal et al., 2012], respectively.

As DECODE is designed to analyze mutations with the same copy number background, we limited the samples to be analyzed by DECODE to those where (a) there are ≥ 50 diploid mutations, and (b) the diploid mutations constitute $\geq 60\%$ of all mutations. Of the 30 retained samples, one (sample P2) contains 123,969 diploid mutations.

Although DECODE successfully deconvoluted this sample, the prohibitive memory demand for the co-clustering matrix (≈ 270 GB) prevented us from scoring **task 2B**. Therefore, P2 was excluded from our analysis, and we reported the results for the remaining 29 synthetic tumors. Following DREAM protocol, subchallenge scores were reported as **NA** for the samples that we excluded.

For each sample, the input data for DECODE consisted of all diploid and heterozygous mutations (major CN = minor CN = 1). Because the DREAM Challenge does not account for the SFS tail, we implemented DECODE without a tail. The result from DECODE thus consisted of cluster count H , the distribution of $\theta_H = \{p_1, \dots, p_H, \omega_1, \dots, \omega_H\}$, and the expected SFS $\left\{ \mathbb{E} \left(S^{[*]}(b) \middle| \theta_{h|H} \right), \text{bin } b = 1, \dots, N_{\text{bins}} \right\}$ for each cluster $h = 1, \dots, H$ with parameters $\theta_{h|H} = \{p_h, \omega_h\}$ under filters $*$ = $\left(L^{[A]}, M^{[A]} \right), \left(L^{[B]}, M^{[B]} \right), \left(L^{[V]}, M^{[V]} \right)$.

Because DECODE’s clusters are ordered such that $p_1 > \dots > p_H$, the cellular prevalence (**task 1A**) corresponds to the truncal cluster’s cancer cell fraction (CCF) $c_1 = 2 \cdot p_1$. The prediction for subclone count (**task 1B**) is H . To score the inference of subclonal cellular prevalences (**task 1C**), DREAM requires the CCF and mutation count for each subclone. Similar to **task 1A**, DECODE predicted the subclonal CCFs to be $2 \cdot p_1, \dots, 2 \cdot p_H$. As the inference A subset has the lowest variant read count threshold ($L^{[A]} < L^{[B]}, L^{[V]}$), we assumed that its deconvoluted SFS is the closest in shape to the unfiltered data, therefore the subclonal mutation counts follow ratio $\omega_1^{[A]} : \dots : \omega_H^{[A]}$.

DREAM’s scores for mutational assignment (**task 2A-B**) concern all mutations in the sample, and not limited to only diploid or heterozygous regions. Therefore, we first implemented GRITIC [Baker et al., 2024] to find the mutation multiplicities. For mutation i with observed VAF f_i , multiplicity m_i and total copy number c_i in a sample with purity ρ , we defined its normalized VAF to be the expected VAF if the mutation were diploid [Tarabichi et al., 2021]:

$$\bar{f}_i = f_i \cdot \frac{\rho}{2\rho + 2(1-\rho)} \bigg/ \frac{m_i \cdot \rho}{c_i \cdot \rho + 2(1-\rho)}$$

We then found the bin b_i that contained \bar{f}_i . Because the SFS from inference A is closest in shape to the raw SFS, we expected the unfiltered mutation counts for different clusters in this bin to follow the ratio $\mathbb{E} \left(S^{[A]}(b_i) \middle| \theta_{1|H} \right) : \dots : \mathbb{E} \left(S^{[A]}(b_i) \middle| \theta_{H|H} \right)$. Therefore, the probability that mutation i to belong to cluster h is

$$\mathbb{P}(i \in h) = \mathbb{E} \left(S^{[A]}(b_i) \middle| \theta_{h|H} \right) \bigg/ \sum_{h'=1, \dots, H} \mathbb{E} \left(S^{[A]}(b_i) \middle| \theta_{h'|H} \right)$$

To determine the mutation’s hard assignment (**task 2A**), we attributed it to the cluster with the highest probability. The co-clustering matrix for **task 2B** consisted of the probabilities that each pair of mutations i and j belong to the same cluster:

$$\mathbb{P}_{i,j} = \sum_{h=1, \dots, H} \mathbb{P}(i \in h) \cdot \mathbb{P}(j \in h)$$

for $i \neq j$ and $\mathbb{P}_{i,j} = 1$ for $i = j$.

DECODE’s predictions for **tasks 1A-2B** were scored against ground truth using the evaluation scripts provided by the DREAM Challenge [Salcedo et al., 2020]. We then compared DECODE’s raw scores against 31 existing methods and five reference algorithms previously analyzed by DREAM’s authors [Salcedo et al., 2025] (Fig. 2 and Supplementary Figs. 1-3). To rank DECODE’s performance against the originally benchmarked algorithms, we computed the normalized score for each method in each task:

$$\bar{s}_{\text{method}} = \frac{s_{\text{method}} - s_{\text{min}}}{s_{\text{max}} - s_{\text{min}}}$$

where s_{method} , s_{max} and s_{min} are the method’s average raw score, and best and worst average scores among all software, respectively.

B.2 Comparison of computational costs

We compared the computational runtime between DECODE, PhyloWGS [Deshwar et al., 2015] and SciClone [Miller et al., 2014] across the 29 synthetic tumors included in our DREAM Challenge study. All methods were executed on the same hardware platform, a Mac mini (2023) with an Apple M2 Pro chip. Each package was executed with default settings. Any run that did not complete within 10^6 seconds (≈ 28 hours) was reported as incomplete.

B.3 Detection and characterization of neutral tail

The ICGA-TCGA DREAM Challenge was not designed to validate the identification and profiling of the neutral tail component. Therefore, we performed a second benchmark study to quantify DECODE’s performance in detecting and characterizing the tail in the SFS.

We sampled each simulation’s purity ρ , total observed mutation count m and coverage distribution φ randomly and independently from real cancer samples in The Cancer Genome Atlas (TCGA, more details below) with mean coverage $\geq 30\times$. Each simulation further assumed H clones with hierarchy $(\sigma_1, \dots, \sigma_H)$, where $\sigma_h \in \{0, \dots, h-1\}$ denotes the clonal parent of subclone h and $\bar{\sigma}_h = \{h, \sigma(h), \sigma(\sigma(h)), \dots\}$ details its clonal lineage. Clone 1 was assumed to be truncal, hence $\sigma_1 = 0$ and $\sigma_2, \dots, \sigma_H \geq 1$. The mutations in subclone h ’s MRCA constituted a fraction $\pi_h^{(m)}$ of the observed data (Fig. 1A), and the tail accounted for the remaining fraction $\pi_0^{(m)}$. We assumed the sample consists of $n_{\text{cancer}} = 10^5$ cancer cells, a fraction $\pi_h^{(c)}$ of which belongs to subclone h . The total cell count is then $n_{\text{total}} = n_{\text{cancer}}/\rho$, contaminated by $n_{\text{total}} - n_{\text{cancer}}$ normal cells, therefore there are $2 \cdot n_{\text{total}}$ allele copies at any genomic site.

Each mutation was located at VAF $f = k/(2 \cdot n_{\text{total}})$ in the true neutral tail with probability $k^{-\alpha} / \sum_{k_*=1}^{n_{\text{cancer}}} k_*^{-\alpha}$, where $k = 1, \dots, n_{\text{cancer}}$ is the number of cells (and hence allele copies) where the mutation is present, and α is the tail power. The mutation was sequenced with r reads with probability $\varphi(r)$, then its variant read count was sampled from $z \sim \text{Binomial}(r, f)$. We filtered out mutations with $z \leq 2$, mimicking typical variant-calling pipelines that exclude mutations supported by little evidence. Using this scheme, the simulator yielded $m \cdot \pi_0^{(m)}$ tail mutations in the observed dataset. The $m \cdot \pi_h^{(m)}$ observed mutations in each cluster h were simulated similarly with true VAF $f = n_h/(2 \cdot n_{\text{total}})$, where $n_h = n_{\text{cancer}} \cdot \sum_{h_*: h \in \bar{\sigma}_{h_*}} \pi_{h_*}^{(c)}$ is the number of cells belonging to clone h and its subclones. The simulated mutational read count table was then analyzed by DECODE and MOBSTER [Caravagna et al., 2020], currently the only alternative mutation clustering algorithm that supports tail detection, with default settings for both methods.

The benchmark consisted of simulations with subclone count $H = 1$ (with hierarchy $\sigma_1 = 0$), $H = 2$ (with $(\sigma_1, \sigma_2) = (0, 1)$) and $H = 3$ (with $(\sigma_1, \sigma_2, \sigma_3) = (0, 1, 1)$), resulting in three simulation batches. Each batch consisted of 1800 simulated samples with tail (T-type) and 200 samples consisting of only mutation clusters without any tail mutations (NT-type). For each T-type simulation, tail power α was sampled from Uniform(0.9, 2.5), and both subclonal cell proportions $(\pi_1^{(c)}, \dots, \pi_H^{(c)})$ and mutation proportions $(\pi_0^{(m)}, \dots, \pi_H^{(m)})$ were sampled from Dirichlet(1, ..., 1), conditioned on $\pi_0^{(m)}, \dots, \pi_H^{(m)} \geq 0.1$, $\pi_1^{(c)}, \dots, \pi_H^{(c)} \geq 0.225$ and $\min_{h, h'} |\pi_h^{(c)} - \pi_{h'}^{(c)}| \geq 0.1$. For NT-type simulations, cell proportions $(\pi_1^{(c)}, \dots, \pi_H^{(c)})$ and mutation proportions $(\pi_1^{(m)}, \dots, \pi_H^{(m)})$ were similarly simulated and conditioned.

We evaluated decomposition results from DECODE and MOBSTER based on their detection and characterization of the tail component. The results were classified as false positive (FP) if they indicated the existence of a tail in NT-type simulations, and false negative (FN) if they failed to detect the tail in T-type samples. True-positive

results were further evaluated based on the accuracy of the inferred tail power and tail mutation proportion, compared to the simulations’ ground-truth values of α and $\pi_0^{(m)}$. Both predictions were readily available from MOBSTER, as was the inferred tail power from DECODE. However, DECODE applied mutation filters before decomposing the SFS (Fig. 1B), hence its mutation counts were not directly comparable to those in the simulated data. DECODE’s decomposition of the *inference* A subsample containing H clusters can be represented as $\{s_h^{[A]}(b)\}$, where $s_h^{[A]}(b)$ is the expected mutation count from compartment h in bin b . The corrected tail mutation proportion from DECODE is then:

$$\hat{\pi}_0^{(m)} = \left(\sum_{b=1}^{N_{\text{bins}}} n_b \cdot \frac{s_0^{[A]}(b)}{\sum_{h=0}^H s_h^{[A]}(b)} \right) / \sum_{b=1}^{N_{\text{bins}}} n_b$$

where n_b is the number of observed mutations in bin b .

We compared MOBSTER’s and DECODE’s accuracy in inferring tail parameter $\theta_* = \alpha, \pi_0^{(m)}$ using root mean square error (RMSE):

$$\text{RMSE}(\theta_*) = \sqrt{\frac{1}{|I|} \sum_{i \in I} (\theta_*^{[i]} - \hat{\theta}_*^{[i]})^2}$$

where $\hat{\theta}_*^{[i]}$ is the prediction compared to the true value $\theta_*^{[i]}$ in simulation i , restricted to the subset I of simulations where the inferences of each method were true-positive.

We considered five confounding factors that may cause DECODE and MOBSTER’s results to be inaccurate: sample purity, coverage depth, mutation count, error in estimating cluster count, and the entropy caused by the overlap between the tail and mutation clusters. The cluster count error was defined as the inferred cluster count from each method minus the true cluster count. To quantify the tail entropy, let $n_h(b)$ be the number of mutations in bin b that belong to compartment h in the ground truth. We defined $p_{\text{tail}}(b) = n_0(b) / \sum_h n_h(b)$ and $p_{\text{clusters}}(b) = 1 - p_{\text{tail}}(b)$ to be the proportions of mutations in bin b that belong in the tail or clusters, respectively. The tail entropy was then based on Shannon’s diversity index:

$$\text{Tail entropy} = - \sum_{b=1}^{N_{\text{bins}}} [p_{\text{tail}}(b) \cdot \log_2(p_{\text{tail}}(b)) + p_{\text{clusters}}(b) \cdot \log_2(p_{\text{clusters}}(b))] \quad (6)$$

using the convention that $0 \cdot \log_2(0) = 0$. Higher tail entropy indicates greater overlap between the tail and mutation clusters, rendering the tail power and mutation count less identifiable (Supplementary Figs. 4-6).

C Analysis of paired AML samples at diagnosis and relapse

Paired acute myeloid leukemia (AML) samples from 11 patients, collected at diagnosis and relapse, together with matched control samples described in [Shlush et al., 2017], were downloaded in BAM format from the EGA repository (ID: EGAD00001003234). The reads had previously been aligned with the hg19 reference genome using NovoAlign. All samples were additionally processed using Picard MarkDuplicates and GATK BaseRecalibrator (v4.1.1.0) [DePristo et al., 2011]. The average effective sequencing coverage was approximately $59\times$. Somatic variants were identified using GATK MuTect2 (v4.1.1.2) and filtered with FilterMutectCalls, incorporating contamination estimates from CalculateContamination and orientation bias statistics from LearnReadOrientationModel. Variants from matched diagnosis and relapse samples were combined using GATK CombineVariants, and variants failing filters in either sample were excluded to avoid inconsistencies caused by independent filtering. Retained variants were annotated using Variant Effect Predictor (VEP v97) [McLaren et al., 2016].

Copy number alterations were identified using GATK (v4.1.1.2) tools [DePristo et al., 2011] with the following tools and non-default parameters: PreprocessIntervals, CollectReadCounts, CreateReadCountPanelOfNormals (minimum-interval-median-percentile: 5.0), DenoiseReadCounts, CollectAllelicCounts, ModelSegments (number-of-smoothing-iterations-per-fit: 1, kernel-variance-copy-ratio: 0.8, kernel-variance-allele-fraction: 0.8, smoothing-credible-interval-threshold-allele-fraction: 10, smoothing-credible-interval-threshold-copy-ratio: 10), and CallCopyRatioSegments. These parameters were adjusted according to the authors’ recommendations for datasets exhibiting excessive numbers of CNV segments caused by noisy segment-specific read counts, as assessed using the GATK visualization tools PlotDenoisedCopyRatios and PlotModeledSegments.

DECODE was later used to analyze the paired samples at diagnosis and relapse ($n = 11$ patients). To remove samples with extremely low purity, we computed $\%_{\text{VAF}>0.2}$ for each patient, the percentage of diploid mutations with $\text{VAF} > 0.2$ at diagnosis or relapse, and excluded those with $\%_{\text{VAF}>0.2} < 5\%$. This eliminated Patient 9 ($\%_{\text{VAF}>0.2} = 2.4\%$) and retained ten patients ($\%_{\text{VAF}>0.2} = 12.5 - 60.6\%$).

For each patient, we implemented three mutational decomposition pipelines: (a) multi-sample SciClone [Miller et al., 2014] using paired data at diagnosis and relapse, (b) MOBSTER [Caravagna et al., 2020] and (c) DECODE, applied individually for each sample. The mutational data analyzed by all pipelines were filtered such that (1) multi-allelic loci were removed, (2) only diploid SNVs were retained, and (3) mutations with $\text{VAF} > 0.75$ were removed, as they likely occurred on regions with loss of heterozygosity (LOH), given that the copy number is 2. For multi-sample SciClone, mutations found in only one sample were given $\text{VAF} = 0$ in the other sample. By default, SciClone only analyzed mutations sequenced at depth $\geq 100\times$ (minimumDepth = 100), which would remove 89–99% of mutations in each sample. Therefore, we applied SciClone on paired data with minimumDepth = 25 and default values for other hyper-parameters. MOBSTER and DECODE were conducted on single samples with default settings.

To illustrate the joint SFS to be expected under our proposed evolution model in Fig. 3I, we created synthetic data using TEMULATOR [Heide, 2022]. The simulation was configured to contain only one clone starting at time $t = 0$, evolving with birth rate 1.5, death rate 0.2 and mutation rate 200. The synthetic diagnosis and relapse samples were taken at time $t = 7$ and $t = 10$, respectively, such that each sample contained 500 cells. We assumed tumor purity = 1 in both simulated samples. Mutations were labeled “clonal” if their CCF $\geq 90\%$ and “tail” otherwise. For each mutation i with CCF c_i in each sample, we first sampled its total read count r_i from the corresponding empirical coverage distribution from Patient 1, then generated its alternative read count as $z_i \sim \text{Binomial}(r_i, c_i/2)$. The joint and marginal SFS were then tallied from mutations with $z_i \geq 3$.

To perform the dN/dS analysis, mutations from diagnosis and relapse were combined for each patient. The patient’s “cluster mutations” and “tail mutations” consisted of variants either classified accordingly in both samples, or assigned accordingly in one sample and undetected in the other. We then implemented R package `dndscv` [Martincorena et al., 2017] to compute the dN/dS ratio for each mutation category and patient.

D Analysis of primary tumors in TCGA

We applied DECODE to tumor samples from The Cancer Genome Atlas (TCGA) [Weinstein et al., 2013]. The cohort consisted of 10,106 samples for which both mutation calls from MuTect2 [Benjamin et al., 2019] and allele-specific copy number profiles from ASCAT3 [Van Loo et al., 2010] were available, across 8,033 patients and 32 cancer types. Our analysis was limited to 4,666 samples with mean coverage $\geq 30\times$ that were predominantly diploid, i.e., with $\geq 60\%$ mutations residing on diploid regions (major CN = minor CN = 1) and ≥ 50 diploid mutations in total.

We examined DECODE’s consistency by analyzing 57 pairs of biological replicates across 49 patients in TCGA. DECODE and SciClone [Miller et al., 2014] were implemented to cluster mutations in individual samples, restricted to diploid mutations as required by both algorithms. Similar to the AML study, SciClone was applied with `minimumDepth = 25`, and default values for other hyper-parameters. DECODE was applied in the default configuration, i.e., the best fits with and without tail were compared to determine the parsimonious deconvolutions. Although DECODE finished all analyses, SciClone did not complete for two patients (TCGA-DJ-A1QF and TCGA-AZ-4315) within 1 day; these patients were thus excluded in SciClone’s results. We compared the assignments of mutations to SFS compartments (i.e., either the tail or specific clusters) across biological replicates from each algorithm to evaluate its consistency. For each patient, we included mutations that were (i) present in both replicates, (ii) located in diploid regions, and (iii) assigned to specific compartments in each replicate. The third condition excluded mutations unassigned by SciClone due to their locations in the SFS being where multiple subclones overlapped. The adjusted Rand index (ARI) for the biological replicate pair is then

$$\text{ARI} = \frac{\sum_{i,j} C(n_{i,j}, 2) - \left[\sum_i C(n_i^{(1)}, 2) \right] \cdot \left[\sum_j C(n_j^{(2)}, 2) \right] / C(n, 2)}{\frac{1}{2} \left[\sum_i C(n_i^{(1)}, 2) + \sum_j C(n_j^{(2)}, 2) \right] - \left[\sum_i C(n_i^{(1)}, 2) \right] \cdot \left[\sum_j C(n_j^{(2)}, 2) \right] / C(n, 2)}$$

where n is the total mutation count, $n_i^{(1)}$ is the number of mutations in compartment i in replicate 1, $n_j^{(2)}$ is the number of mutations in compartment j in replicate 2, $n_{i,j}$ is the number of mutations assigned to compartment i in replicate 1 and compartment j in replicate 2, and $C(n, 2) = \binom{n}{2}$. ARI scores for pairs of biological replicates were computed with R package `mc1ust` [Scrucca et al., 2023], then compared between DECODE and SciClone.

We also evaluated the consistency in DECODE-inferred tail powers. Among the 49 patients with multi-region WGS data, DECODE detected the tail in more than one region in 29 patients, resulting in 35 pairs of biological replicates where tail powers were assigned for both samples. For each pair of same-patient samples (i, j) with tail powers α_i and α_j , we defined the pairwise α difference as $d(\alpha_i, \alpha_j) = |\alpha_i - \alpha_j|$. We then performed the permutation test, where the observed median same-patient α difference was compared to the medians of $d(\alpha_i, \alpha_j)$ where i and j were randomly permuted among samples from all patients. The reported p-value reflected the proportion of 10,000 permutations that resulted in lower medians than the true median same-patient difference.

We next examined DECODE’s results for the whole diploid TCGA dataset. To evaluate the relationship between patient characteristics and outcome on one hand and DECODE-inferred tail power and clonality at diagnosis on the other, we restricted the analysis to one sample per patient, selecting the sample with highest coverage when biological or technical replicates were available. Our analysis then consisted of $n = 4,204$ unique tumors across 32 cancer types. For each sample, the input for DECODE likewise consisted only of diploid mutations.

We compared sample purities predicted from DECODE against the values inferred previously from ASCAT3. As in the DREAM Challenge study, the purities can be estimated from truncal cluster VAFs. However, inaccurate copy number inference can produce spurious subclones with $\text{VAF} \gg 0.5$ (i.e., $\text{CCF} \gg 1$), for instance due to genomic regions with (major CN, minor CN) = (2, 0) being miscalled as (1, 1). Therefore, DECODE’s inferred sample purity was computed as $2 \times \max \{p_h : p_h \leq 0.55, h = 1, \dots, H\}$. On the other hand, ASCAT3 depends on the genomic regions with diverse copy number states to estimate purity. As a result, purity estimates were unreliable for samples with low aneuploidy (for these cases, ASCAT3 defaulted to purity = 1). We therefore restricted the comparison of purity estimates between DECODE and ASCAT3 to samples in which 10% of the genome or more was aneuploid ($n = 2,333$).

To standardize our comparison of tumor characteristics, DECODE deconvolutions and patient outcomes, we limited the analysis to de novo cancers (4,058 unique tumors across 32 types) and harmonized AJCC pathologic

stage and tumor grade annotations into broader categories. AJCC stage annotations were mapped as follows: stages 0, 0a, 0is and IS were grouped as stage 0; stages I, IA, IB, IB1 as stage I; stages II, IIA, IIB and IIC as stage II; stages III, IIIA, IIIB and IIIC as stage III; and stages IV, IVA, IVB, IVC and X as stage IV; with missing or other annotations being grouped as unknown. Two cases warrant note: stage IS, observed in one TGCT patient (TCGA-2G-AAF1), was grouped into stage 0 following clinical staging convention, and stage X, observed in four BRCA patients, was grouped with stage IV. Tumor grade was primarily grouped into G1, G2, G3 and G4. Grades GB, observed in one OV patient (TCGA-24-2038), and GX, observed in 23 patients across CESC, ESCA, HNSC, KIRC, PAAD, and STAD cancer types, were treated as unknown. For BLCA, where tumors were annotated as low grade or high grade, these categories were retained as reported; high grade was also observed in 2 of 294 UCEC patients and remained as such.

We examined which patient subsets exhibited an association between DECODE-inferred clonality and treatment outcome. We stratified TCGA primary cancers into cohorts by (i) cancer type, (ii) cancer type and AJCC stage, or (iii) cancer type and tumor grade. Samples in each cohort were partitioned into two subgroups by comparing their DECODE-derived cluster counts to the cohort median. Survival analysis was performed only for cohorts in which (a) both subgroups contained ≥ 20 patients, and (b) each subgroup included ≥ 5 death events. A total of 26 cohorts met these criteria: 15 were defined by cancer type alone, three by cancer type and AJCC stage, and eight by cancer type and tumor grade. For each cohort, the hazard ratio (HR) was computed from a Cox proportional hazards model with subgroup identity as the predictor of patient survival. Differences in survival between the two subgroups were evaluated using the log-rank test applied to Kaplan–Meier survival curves. The resulting p-values were adjusted for multiple testing using the Benjamini–Hochberg (BH) procedure, yielding BH-adjusted p-values p_{adj} .

We investigated two TCGA cohorts where partitioning by median cluster count led to statistically significant differences in survival: LGG samples, and LGG grade 3 tumors, both with median cluster count = 1. We first compared tail powers between the partitioned subgroups across both cohorts, limited to samples where DECODE detected a tail (LGG: $n = 301$, LGG grade 3: $n = 127$). We then classified an LGG sample as IDH mutant if it contained one or more mutations that (i) affected IDH1 or IDH2, (ii) passed MuTect2 filtering, and (iii) had moderate or high predicted functional impact (LGG: $n = 231$, LGG grade 3: $n = 87$). We compared the proportion of samples where DECODE detected multiple subclones in each cohort, partitioned by LGG mutant and LGG wild-type, using two-sided Fisher’s exact test.

Code availability

DECODE is available at <https://github.com/dinhngockhanh/DECODE>.

Acknowledgments

YC, MK and KND were supported by the National Cancer Institute of the National Institutes of Health under Award Number R01CA310281. RJ, MKK and MK were supported by the Polish National Science Centre under Award Number UMO-2021/41/B/NZ2/04134. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Competing interests

The authors declare no competing interests.

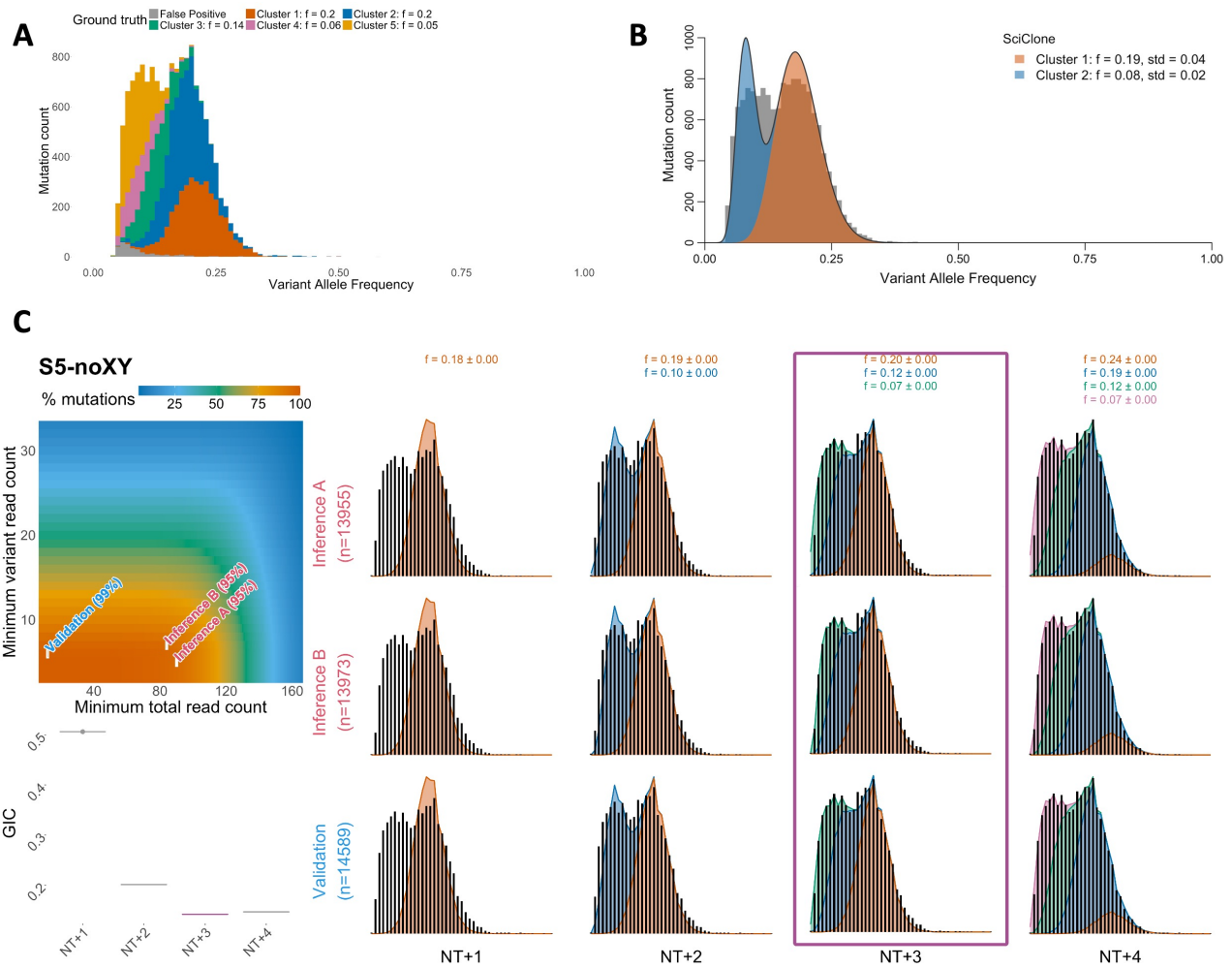
References

- N. Andor, T. A. Graham, M. Jansen, L. C. Xia, C. A. Aktipis, C. Petritsch, H. P. Ji, and C. C. Maley. Pan-cancer analysis of the extent and consequences of intratumor heterogeneity. *Nature Medicine*, 22(1):105–113, 2016.
- T. M. Baker, S. Lai, A. R. Lynch, T. Lesluyes, H. Yan, H. A. Ogilvie, A. Verfaillie, S. Dentre, A. L. Bowes, N. Pillay, et al. The history of chromosomal instability in genome-doubled tumors. *Cancer Discovery*, 14(10):1810–1822, 2024.
- M. A. Beaumont, J.-M. Cornuet, J.-M. Marin, and C. P. Robert. Adaptive approximate Bayesian computation. *Biometrika*, 96:983–990, 2009.
- D. Benjamin, T. Sato, K. Cibulskis, G. Getz, C. Stewart, and L. Lichtenstein. Calling somatic SNVs and indels with Mutect2. *bioRxiv*, page 861054, 2019.
- G. Caravagna, T. Heide, M. J. Williams, L. Zapata, D. Nichol, K. Chkhaidze, W. Cross, G. D. Cresswell, B. Werner, A. Acar, et al. Subclonal reconstruction of tumors by using machine learning and population genetics. *Nature Genetics*, 52(9):898–907, 2020.
- S. L. Carter, K. Cibulskis, E. Helman, A. McKenna, H. Shen, T. Zack, P. W. Laird, R. C. Onofrio, W. Winckler, B. A. Weir, et al. Absolute quantification of somatic DNA alterations in human cancer. *Nature Biotechnology*, 30(5):413–421, 2012.
- D. R. Caswell and C. Swanton. The role of tumour heterogeneity and clonal cooperativity in metastasis, immune evasion and clinical outcome. *BMC Medicine*, 15(1):133, 2017.
- D. Čevič, L. Michel, J. Näf, P. Bühlmann, and N. Meinshausen. Distributional Random Forests: Heterogeneity adjustment and multivariate distributional regression. *Journal of Machine Learning Research*, 23:1–79, 2022.
- K. Cibulskis, M. S. Lawrence, S. L. Carter, A. Sivachenko, D. Jaffe, C. Sougnez, S. Gabriel, M. Meyerson, E. S. Lander, and G. Getz. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nature Biotechnology*, 31(3):213–219, 2013.
- I. Dągogo-Jack and A. T. Shaw. Tumour heterogeneity and resistance to cancer therapies. *Nature Reviews Clinical Oncology*, 15(2):81–94, 2018.
- M. A. DePristo, E. Banks, R. Poplin, K. V. Garimella, J. R. Maguire, C. Hartl, A. A. Philippakis, G. del Angel, M. A. Rivas, M. Hanna, A. McKenna, T. J. Fennell, A. M. Kernytzky, A. Y. Sivachenko, K. Cibulskis, S. B. Gabriel, D. Altshuler, and M. J. Daly. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*, 43(5):491–498, 2011.
- A. G. Deshwar, S. Vembu, C. K. Yung, G. H. Jang, L. Stein, and Q. Morris. PhyloWGS: reconstructing subclonal composition and evolution from whole-genome sequencing of tumors. *Genome Biology*, 16(1):35, 2015.

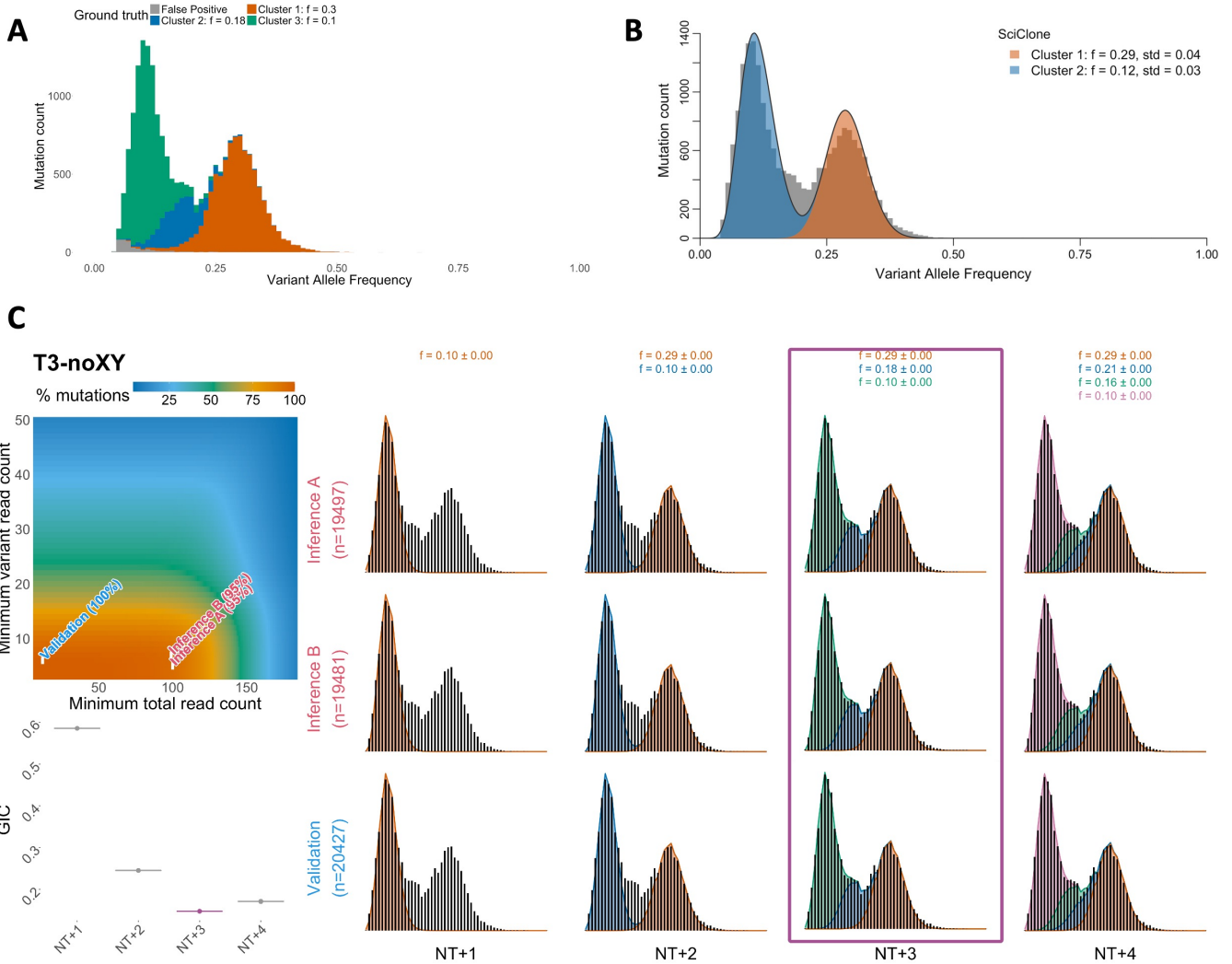
- K. N. Dinh, R. Jaksik, M. Kimmel, A. Lambert, and S. Tavaré. Statistical inference for the evolutionary history of cancer genomes. *Statistical Science*, 35(1):129–144, 2020.
- K. N. Dinh, C. Liu, Z. Xiang, Z. Liu, and S. Tavaré. Approximate Bayesian computation sequential Monte Carlo via random forests. *Statistics and Computing*, 35(6):219, 2025.
- R. Durrett. Population genetics of neutral mutations in exponentially growing cancer cell populations. *The Annals of Applied Probability*, 23(1):230–250, 2013.
- A. D. Ewing, K. E. Houlahan, Y. Hu, K. Ellrott, C. Caloian, T. N. Yamaguchi, J. C. Bare, C. P’ng, D. Waggott, V. Y. Sabelnykova, et al. Combining tumor genome simulation with crowdsourcing to benchmark somatic single-nucleotide-variant detection. *Nature Methods*, 12(7):623–630, 2015.
- C. Ferté, M. Fernandez, A. Hollebecque, S. Koscielny, A. Levy, C. Massard, R. Balheda, B. Bot, C. Gomez-Roca, C. Dromain, et al. Tumor growth rate is an early indicator of antitumor drug activity in phase i clinical trials. *Clinical Cancer Research*, 20(1):246–252, 2014.
- R. C. Griffiths and S. Tavaré. The age of a mutation in a general coalescent tree. *Stochastic Models*, 14(1-2): 273–295, 1998.
- T. Heide. *TEMULATOR: Creation of synthetic tumour sequencing data*, 2022. R package version 1.0.
- S. Konishi and G. Kitagawa. Generalised information criteria in model selection. *Biometrika*, 83(4):875–890, 1996.
- A. B. Krøigård, M. Thomassen, A.-V. Lænkholm, T. A. Kruse, and M. J. Larsen. Evaluation of nine somatic variant callers for detection of somatic mutations in exome and targeted deep sequencing data. *PLOS One*, 11(3):e0151664, 2016.
- I. Martincorena, K. M. Raine, M. Gerstung, K. J. Dawson, K. Haase, P. Van Loo, H. Davies, M. R. Stratton, and P. J. Campbell. Universal patterns of selection in cancer and somatic tissues. *Cell*, 171(5):1029–1041, 2017.
- N. McGranahan and C. Swanton. Clonal heterogeneity and tumor evolution: past, present, and the future. *Cell*, 168(4):613–628, 2017.
- W. McLaren, L. Gil, S. E. Hunt, H. S. Riat, G. R. S. Ritchie, A. Thormann, P. Flicek, and F. Cunningham. The Ensembl Variant Effect Predictor. *Genome Biology*, 17(1):122, 2016.
- C. A. Miller, B. S. White, N. D. Dees, M. Griffith, J. S. Welch, O. L. Griffith, R. Vij, M. H. Tomasson, T. A. Graubert, M. J. Walter, et al. SciClone: inferring clonal architecture and tracking the spatial and temporal patterns of tumor evolution. *PLOS Computational Biology*, 10(8):e1003665, 2014.
- S. Nik-Zainal, P. Van Loo, D. C. Wedge, L. B. Alexandrov, C. D. Greenman, K. W. Lau, K. Raine, D. Jones, J. Marshall, M. Ramakrishna, et al. The life history of 21 breast cancers. *Cell*, 149(5):994–1007, 2012.
- A. Roth, J. Khattra, D. Yap, A. Wan, E. Laks, J. Biele, G. Ha, S. Aparicio, A. Bouchard-Côté, and S. P. Shah. PyClone: statistical inference of clonal population structure in cancer. *Nature Methods*, 11(4):396–398, 2014.
- A. Salcedo, M. Tarabichi, S. M. G. Espiritu, A. G. Deshwar, M. David, N. M. Wilson, S. Dentro, J. A. Wintersinger, L. Y. Liu, M. Ko, et al. A community effort to create standards for evaluating tumor subclonal reconstruction. *Nature Biotechnology*, 38(1):97–107, 2020.

- A. Salcedo, M. Tarabichi, A. Buchanan, S. M. Espiritu, H. Zhang, K. Zhu, T.-H. Ou Yang, I. Leshchiner, D. Anastassiou, Y. Guan, et al. Crowd-sourced benchmarking of single-sample tumor subclonal reconstruction. *Nature Biotechnology*, 43(4):581–592, 2025.
- L. Scrucca, C. Fraley, T. B. Murphy, and A. E. Raftery. *Model-based clustering, classification, and density estimation using mclust in R*. Chapman and Hall/CRC, 2023.
- L. I. Shlush, A. Mitchell, L. Heisler, S. Abelson, S. W. Ng, A. Trotman-Grant, J. J. Medeiros, A. Rao-Bhatia, I. Jaciw-Zurakowsky, R. Marke, et al. Tracing the origins of relapse in acute myeloid leukaemia to stem cells. *Nature*, 547(7661):104–108, 2017.
- M. Sirenko, S. Lee, Z. Sun, R. Chaligne, S. Loghavi, G. Asimomitis, C. K. Brierley, E. Bernard, S. F. Cai, R. M. Myers, et al. Deconvoluting clonal and cellular architecture in IDH-mutant acute myeloid leukemia. *Cell Stem Cell*, 32(7):1102–1121, 2025.
- W. D. Stein, W. D. Figg, W. Dahut, A. D. Stein, M. B. Hoshen, D. Price, S. E. Bates, and T. Fojo. Tumor growth rates derived from data for patients in a clinical trial correlate strongly with patient survival: a novel strategy for evaluation of clinical trial data. *The Oncologist*, 13(10):1046–1054, 2008.
- K. R. Swanson, R. C. Rostomily, and E. Alvord. A mathematical modelling tool for predicting survival of individual patients following resection of glioblastoma: a proof of principle. *British Journal of Cancer*, 98(1):113–119, 2008.
- M. Tarabichi, A. Salcedo, A. G. Deshwar, M. Ni Leathlobhair, J. Wintersinger, D. C. Wedge, P. Van Loo, Q. D. Morris, and P. C. Boutros. A practical guide to cancer subclonal reconstruction from DNA sequencing. *Nature Methods*, 18(2):144–155, 2021.
- The Cancer Genome Atlas Research Network. Comprehensive, integrative genomic analysis of diffuse lower-grade gliomas. *New England Journal of Medicine*, 372(26):2481–2498, 2015.
- S. Turajlic and C. Swanton. Metastasis as an evolutionary process. *Science*, 352(6282):169–175, 2016.
- M. J. van den Bent, P. J. French, D. Brat, J. C. Tonn, M. Touat, B. M. Ellingson, R. J. Young, J. Pallud, A. von Deimling, F. Sahm, D. Figarella Branger, R. Y. Huang, M. Weller, I. K. Mellingerhoff, T. F. Cloughesy, J. T. Huse, K. Aldape, G. Reifenberger, G. Youssef, P. Karschnia, H. Noushmehr, K. B. Peters, F. Ducray, M. Preusser, and P. Y. Wen. The biological significance of tumor grade, age, enhancement, and extent of resection in IDH-mutant gliomas: How should they inform treatment decisions in the era of IDH inhibitors? *Neuro-Oncology*, 26(10):1805–1822, 2024.
- P. Van Loo, S. H. Nordgard, O. C. Lingjærde, H. G. Russnes, I. H. Rye, W. Sun, V. J. Weigman, P. Marynen, A. Zetterberg, B. Naume, et al. Allele-specific copy number analysis of tumors. *Proceedings of the National Academy of Sciences*, 107(39):16910–16915, 2010.
- G. Watterson. On the number of segregating sites in genetical models without recombination. *Theoretical Population Biology*, 7(2):256–276, 1975.
- J. N. Weinstein, E. A. Collisson, G. B. Mills, K. R. Shaw, B. A. Ozenberger, K. Ellrott, I. Shmulevich, C. Sander, and J. M. Stuart. The cancer genome atlas pan-cancer analysis project. *Nature Genetics*, 45(10):1113–1120, 2013.

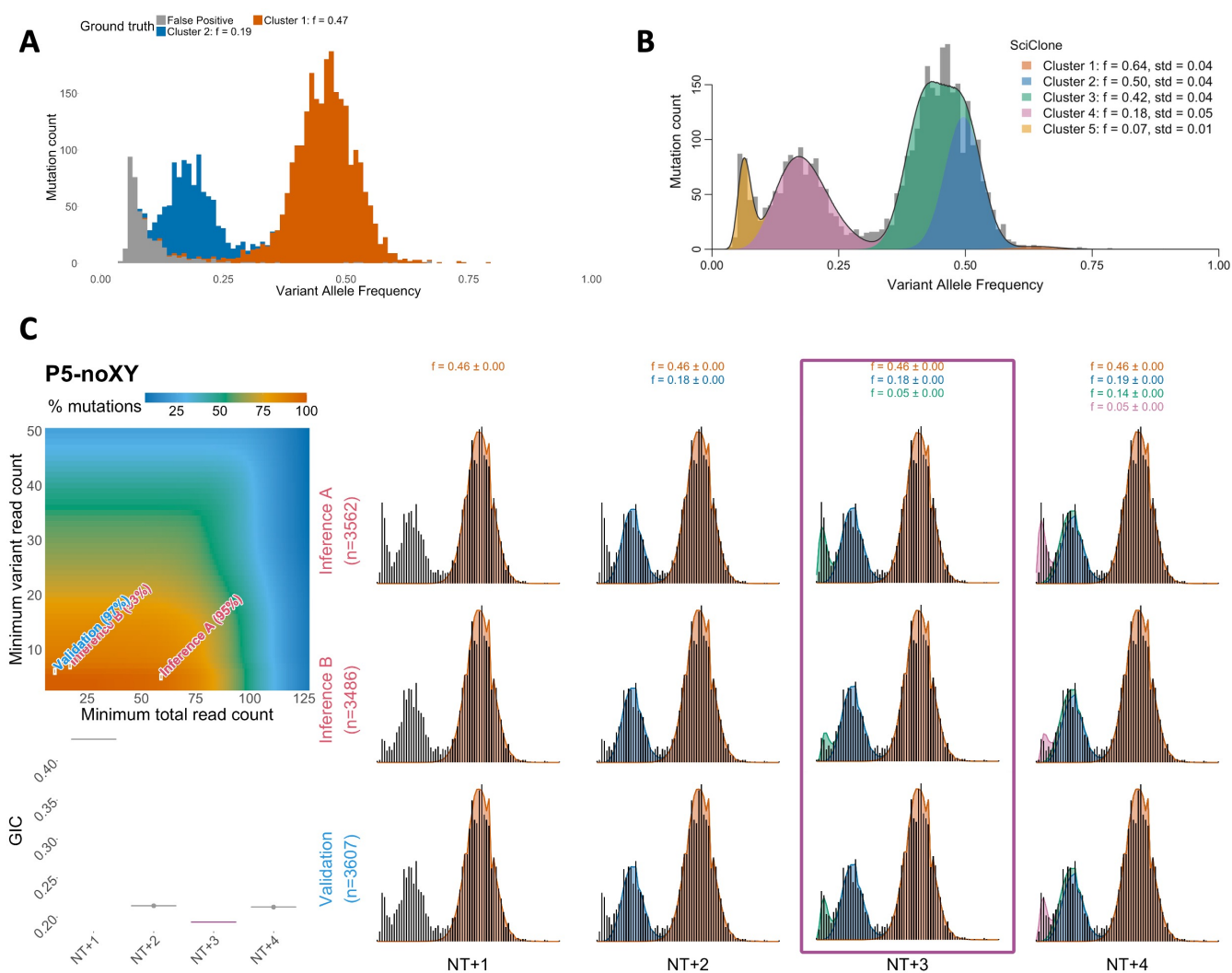
- M. J. Williams, B. Werner, C. P. Barnes, T. A. Graham, and A. Sottoriva. Identification of neutral tumor evolution across cancer types. *Nature Genetics*, 48(3):238–244, 2016.
- X. Xiang, Z. Liu, C. Zhang, Z. Li, J. Gao, C. Zhang, Q. Cao, J. Cheng, H. Liu, D. Chen, et al. IDH mutation subgroup status associates with intratumor heterogeneity and the tumor microenvironment in intrahepatic cholangiocarcinoma. *Advanced Science*, 8(17):2101230, 2021.



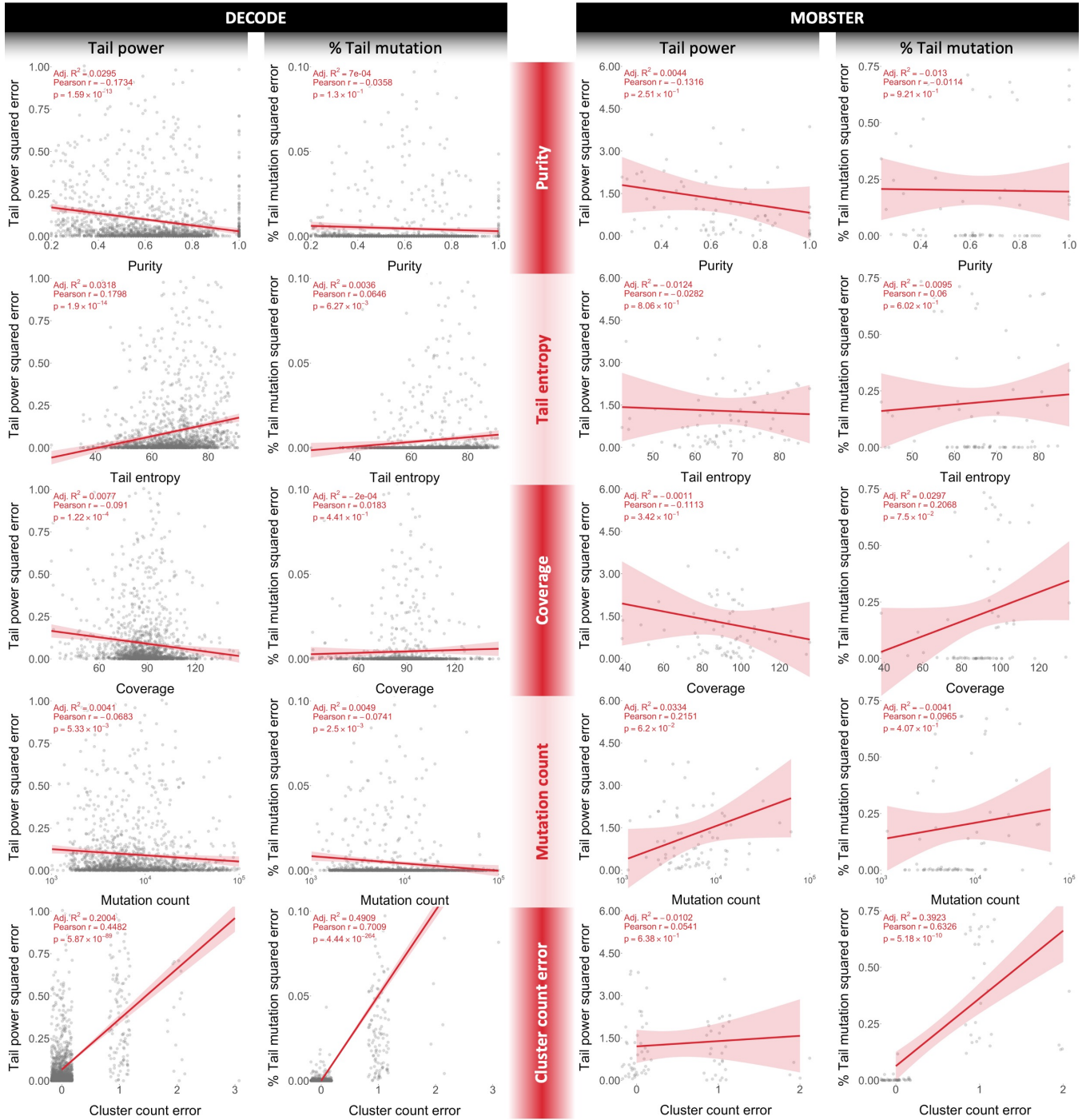
Supplementary Figure 1: Nonexistent clonal cluster results in incorrectly inferred sample purity. **A-B**: Ground truth SFS (**A**) and SciClone's decomposition (**B**) for diploid mutations in DREAM synthetic tumor S5. **C**: DECODE's decomposition of tumor S5. Top left: heatmap of mutation retention frequency for different variant and total read count thresholds. Filtering thresholds for *inference A*, *inference B* and *validation* subsamples are highlighted. Right: Optimal decompositions given cluster count H under assumption of no tail (column $\text{NT}+H$). For each H , the expected observed SFS from the posterior distribution for parameters θ_H (shaded areas, colors denoting clusters) for each subsample is compared against the empirical SFS (black histogram). Means and standard deviations of cluster VAFs are reported on top. Bottom left: Comparison of Generalized Information Criteria (GIC) between different cluster count assumptions. The fit with lowest GIC ($\text{NT}+3$) is chosen.



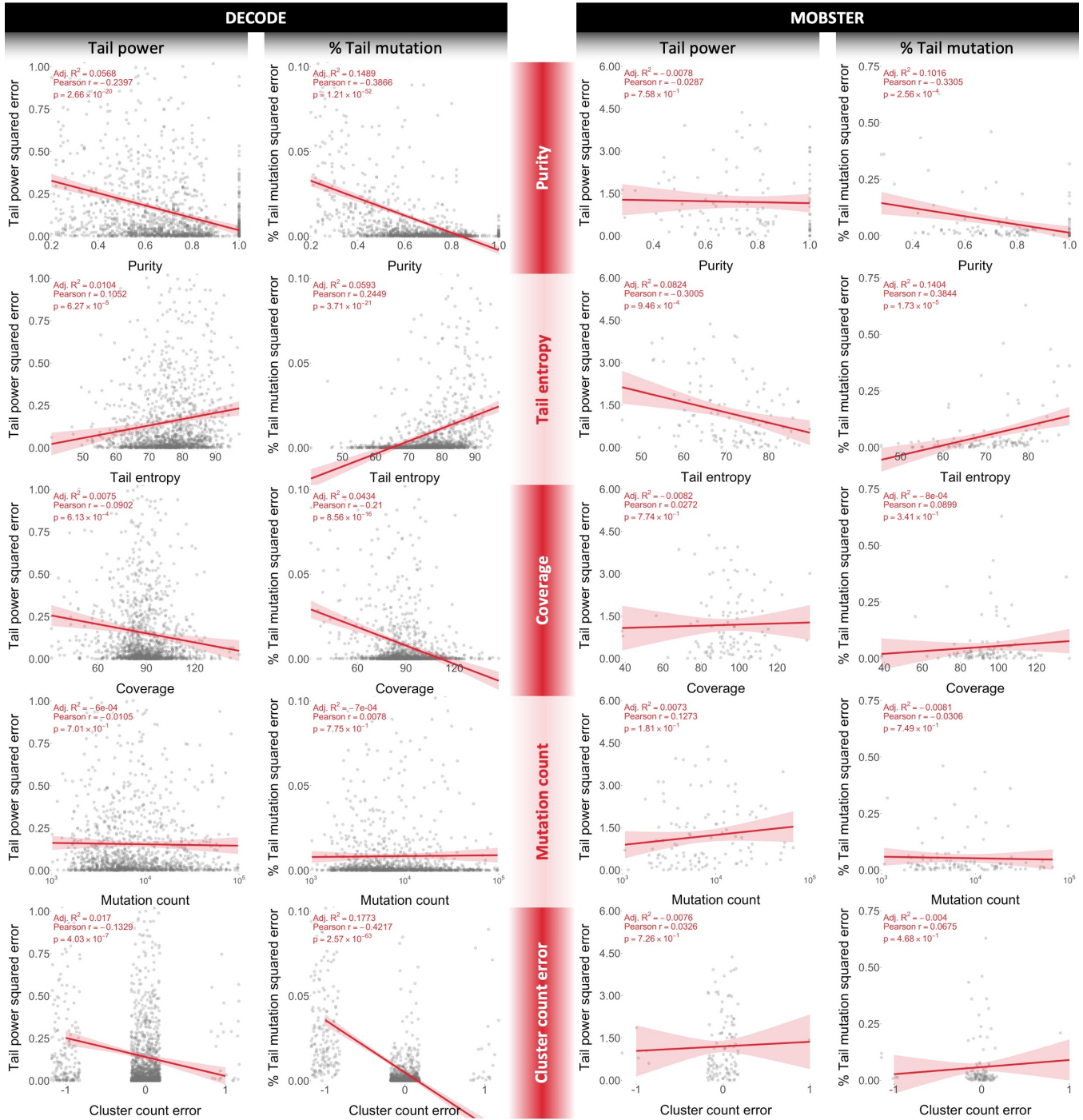
Supplementary Figure 2: DECODE accurately identifies obscured clusters. **A-B**: Ground truth SFS (**A**) and SciClone's decomposition (**B**) for diploid mutations in DREAM synthetic tumor T3. **C**: DECODE's decomposition of tumor T3. Top left: heatmap of mutation retention frequency for different variant and total read count thresholds. Filtering thresholds for *inference A*, *inference B* and *validation* subsamples are highlighted. Right: Optimal decompositions given cluster count H under assumption of no tail (column NT+ H). For each H , the expected observed SFS from the posterior distribution for parameters θ_H (shaded areas, colors denoting clusters) for each subsample is compared against the empirical SFS (black histogram). Means and standard deviations of cluster VAFs are reported on top. Bottom left: Comparison of Generalized Information Criteria (GIC) between different cluster count assumptions. The fit with lowest GIC (NT+3) is chosen.



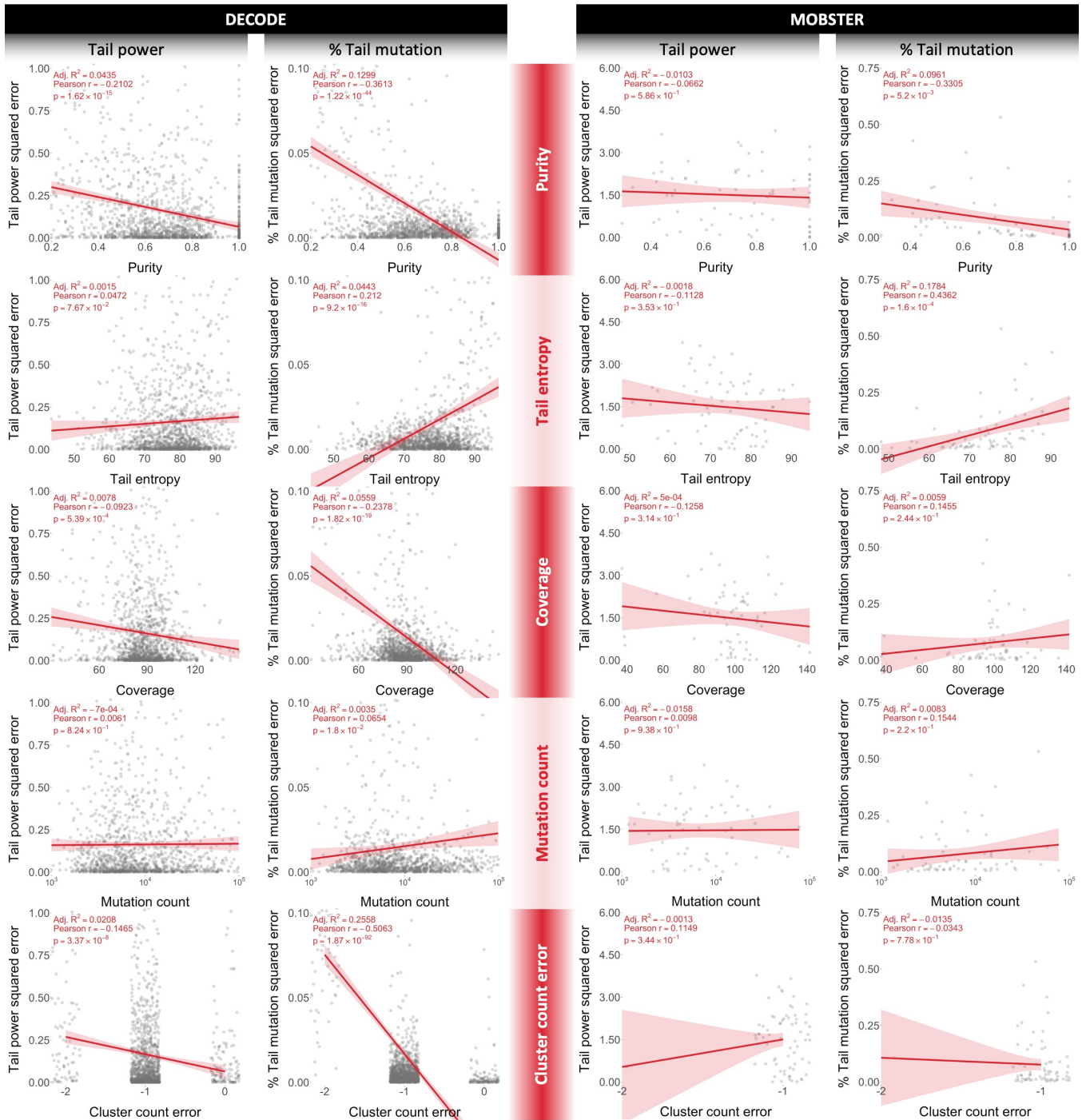
Supplementary Figure 3: DECODE mitigates cluster overfitting. **A-B**: Ground truth SFS (**A**) and SciClone's decomposition (**B**) for diploid mutations in DREAM synthetic tumor P5. **C**: DECODE's decomposition of tumor P5. Top left: heatmap of mutation retention frequency for different variant and total read count thresholds. Filtering thresholds for *inference A*, *inference B* and *validation* subsamples are highlighted. Right: Optimal decompositions given cluster count H under assumption of no tail (column NT+ H). For each H , the expected observed SFS from the posterior distribution for parameters θ_H (shaded areas, colors denoting clusters) for each subsample is compared against the empirical SFS (black histogram). Means and standard deviations of cluster VAFs are reported on top. Bottom left: Comparison of Generalized Information Criteria (GIC) between different cluster count assumptions. The fit with lowest GIC (NT+3) is chosen.



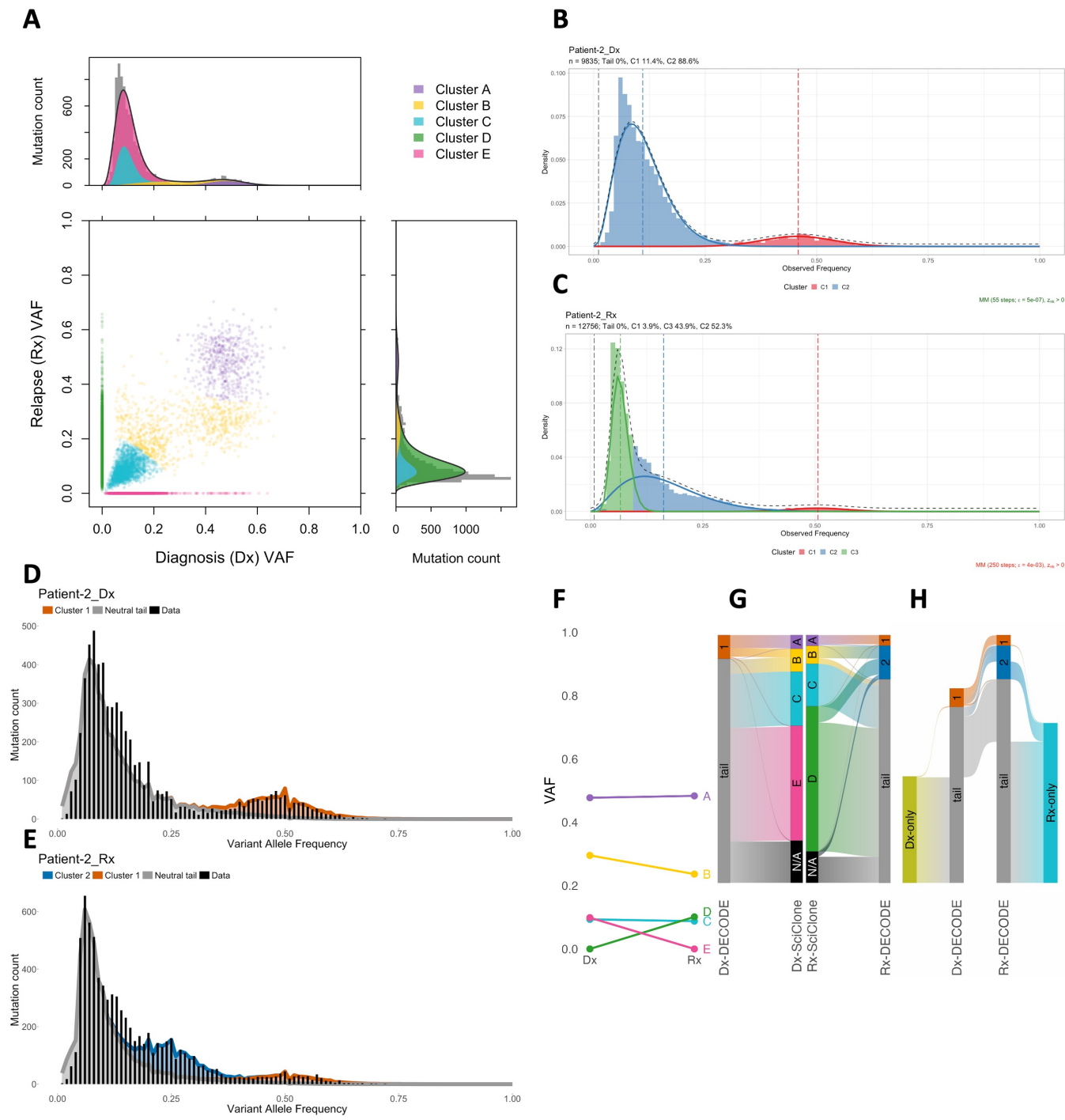
Supplementary Figure 4: Analysis of DECODE's and MOBSTER's errors in estimating tail power and mutation proportion from synthetic samples with true cluster count $H = 1$. Plots display each method's squared errors from individual simulations (dots) as a function of sample purity, tail entropy, coverage depth, observed mutation count, and error in estimating cluster count. Synthetic samples' purity, coverage and mutation count are drawn randomly from TCGA samples. Tail entropy is calculated using Eq. (6). Cluster count error equals the inferred cluster count minus the true value. Red line indicates the linear regression fit, annotations report Pearson's correlation coefficient, its p-value, and the adjusted R^2 .



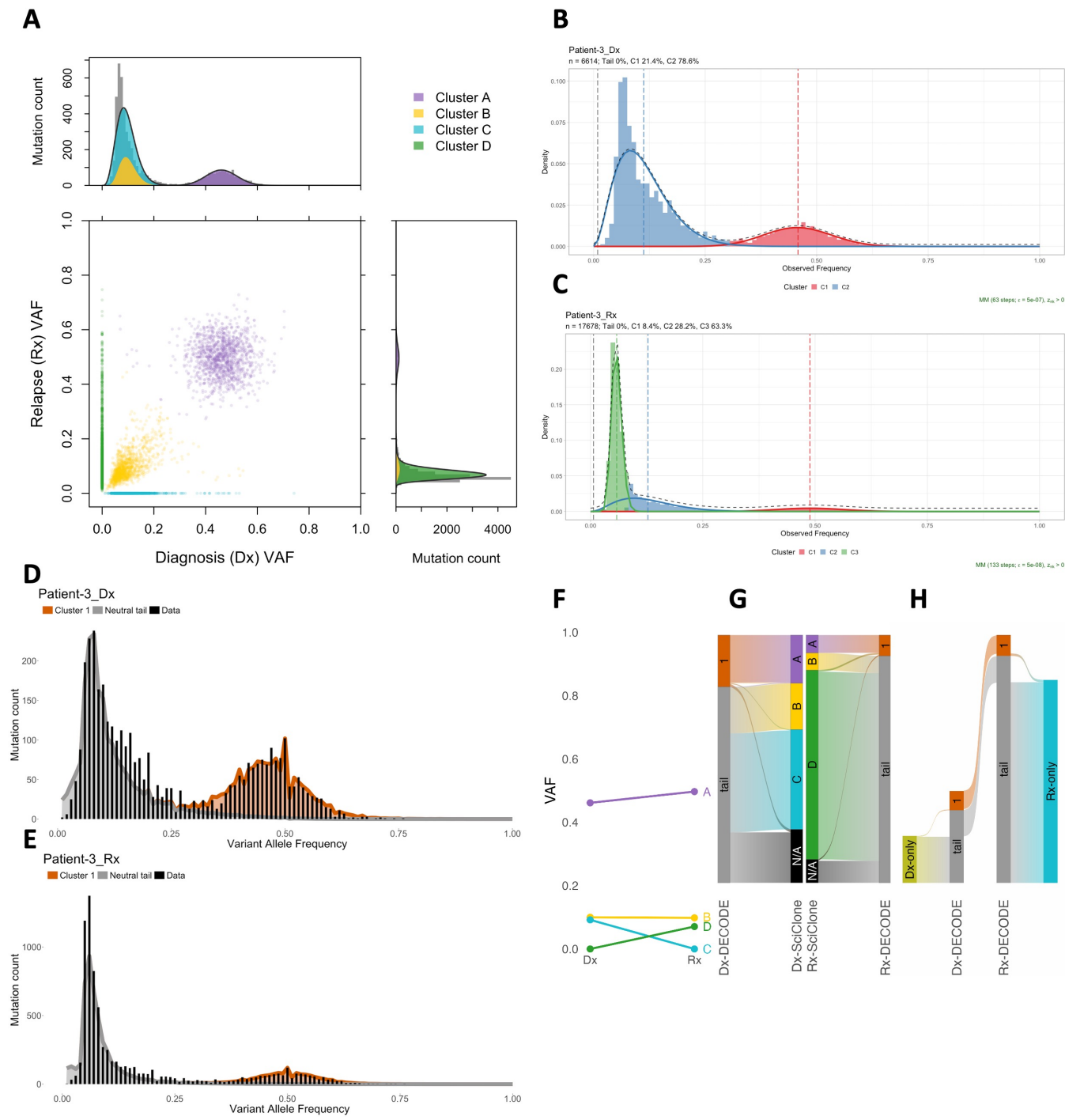
Supplementary Figure 5: Analysis of DECODE's and MOBSTER's errors in estimating tail power and mutation proportion from synthetic samples with true cluster count $H = 2$. Plots display each method's squared errors from individual simulations (dots) as a function of sample purity, tail entropy, coverage depth, observed mutation count, and error in estimating cluster count. Synthetic samples' purity, coverage and mutation count are drawn randomly from TCGA samples. Tail entropy is calculated using Eq. (6). Cluster count error equals the inferred cluster count minus the true value. Red line indicates the linear regression fit, annotations report Pearson's correlation coefficient, its p-value, and the adjusted R^2 .



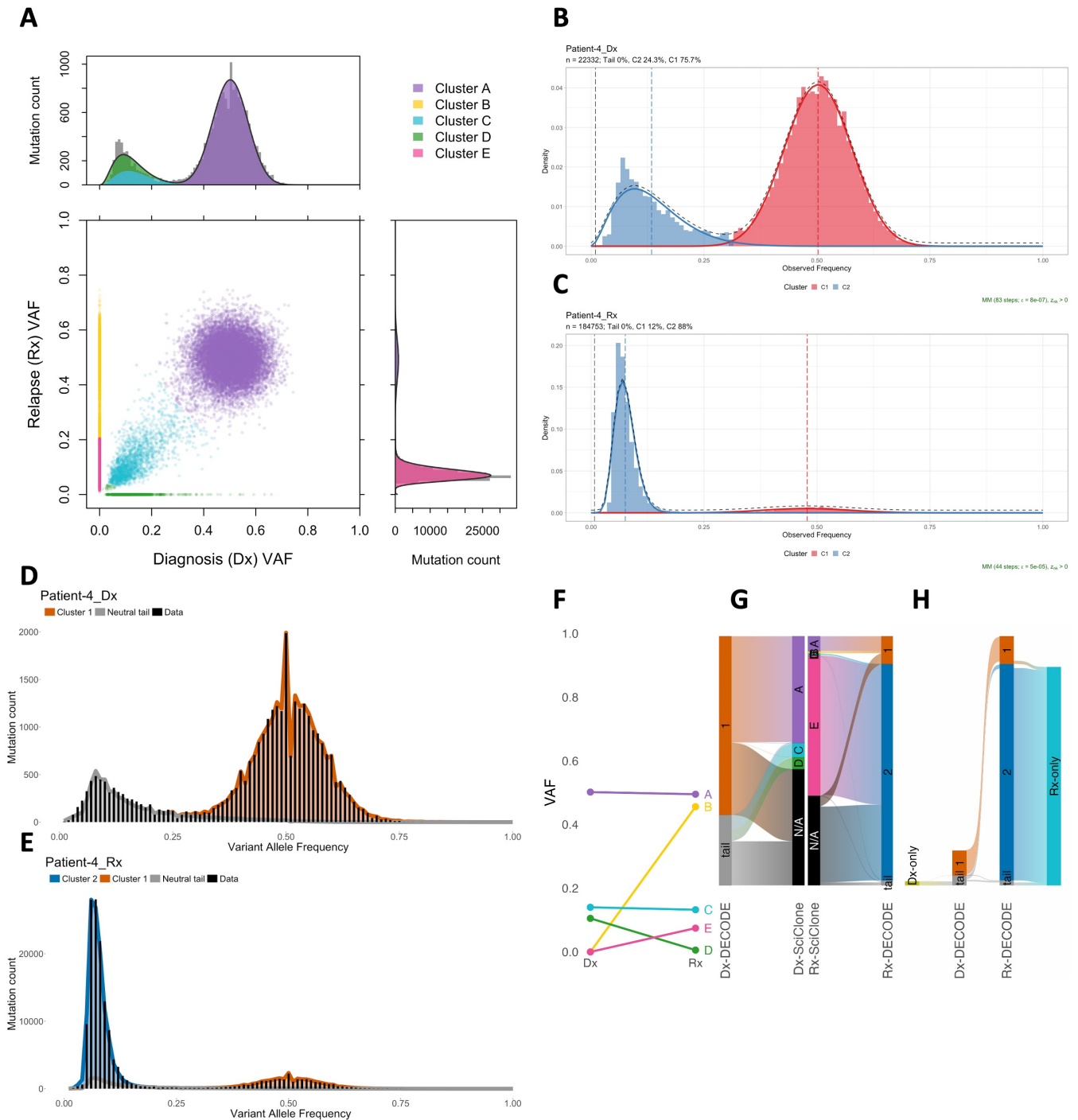
Supplementary Figure 6: Analysis of DECODE's and MOBSTER's errors in estimating tail power and mutation proportion from synthetic samples with true cluster count $H = 3$. Plots display each method's squared errors from individual simulations (dots) as a function of sample purity, tail entropy, coverage depth, observed mutation count, and error in estimating cluster count. Synthetic samples' purity, coverage and mutation count are drawn randomly from TCGA samples. Tail entropy is calculated using Eq. (6). Cluster count error equals the inferred cluster count minus the true value. Red line indicates the linear regression fit, annotations report Pearson's correlation coefficient, its p-value, and the adjusted R^2 .



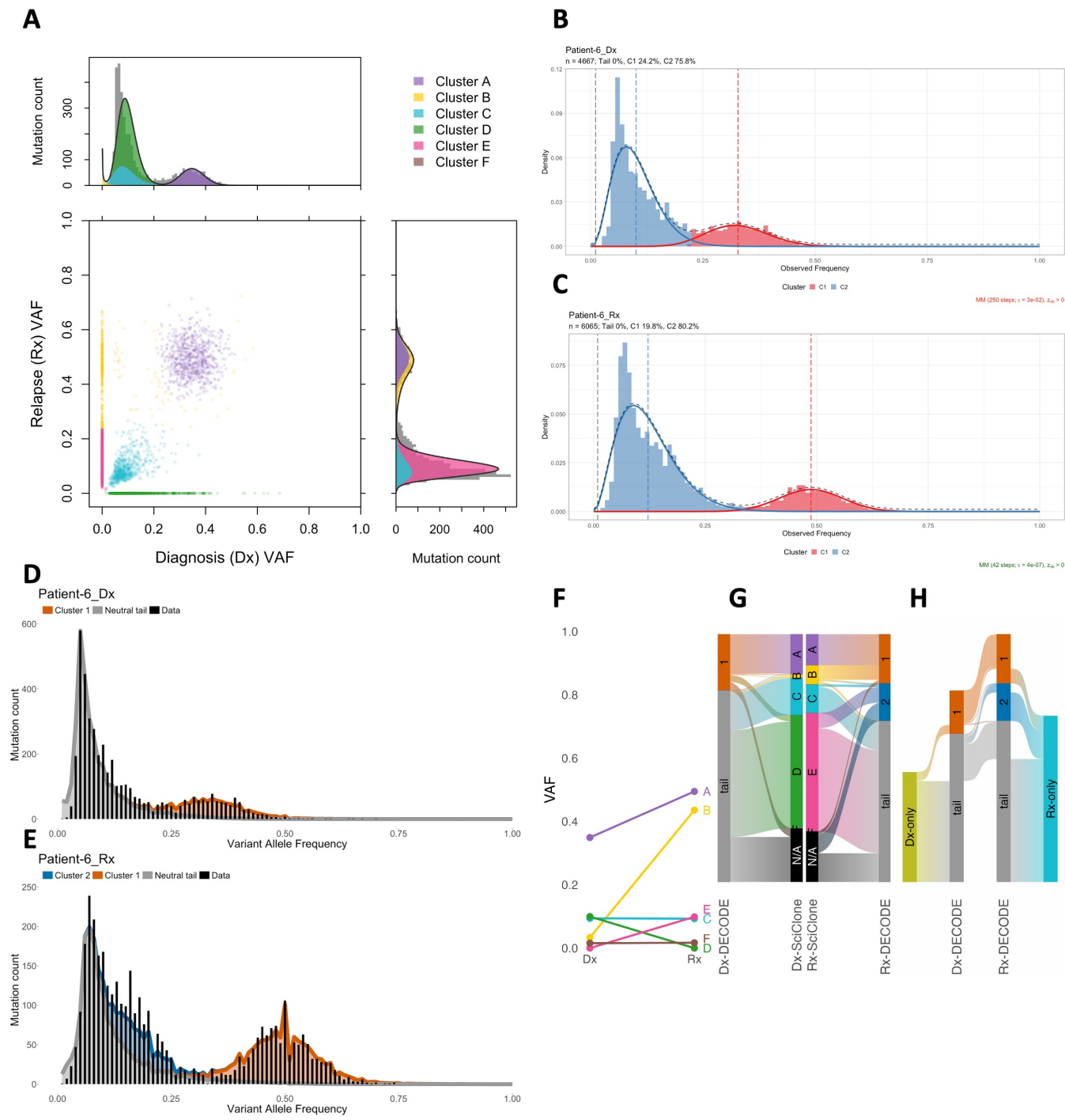
Supplementary Figure 7: Mutation clustering for patient AML-2. **A**: SciClone's joint clustering of mutations at diagnosis and relapse. **B-E**: Deconvolution of mutations using MOBSTER (**B**, **C**) and DECODE (**D**, **E**) for each sample at diagnosis (**B**, **D**) and relapse (**C**, **E**). **F**: Mean VAFs between diagnosis and relapse of subclones inferred from SciClone. **G**: Alluvial diagrams of subclonal assignments for mutations at diagnosis (left) and relapse (right) between DECODE and SciClone. Widths are proportional to mutation fractions. **H**: Alluvial diagram of DECODE subclonal assignments for mutations at diagnosis and relapse. Outer bars contain mutations unique to each sample.



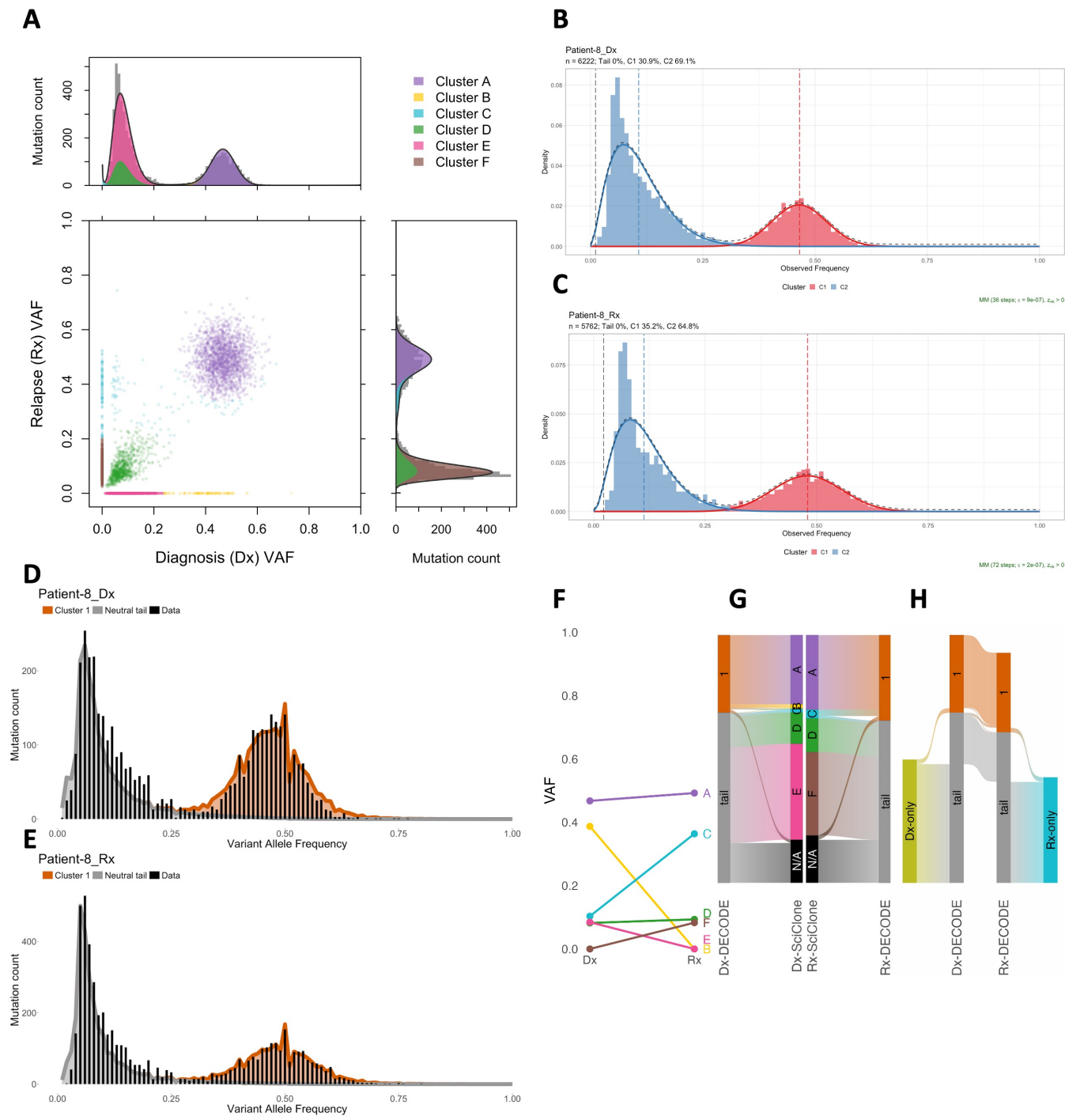
Supplementary Figure 8: Mutation clustering for patient AML-3. **A**: SciClone's joint clustering of mutations at diagnosis and relapse. **B-E**: Deconvolution of mutations using MOBSTER (**B**, **C**) and DECODE (**D**, **E**) for each sample at diagnosis (**B**, **D**) and relapse (**C**, **E**). **F**: Mean VAFs between diagnosis and relapse of subclones inferred from SciClone. **G**: Alluvial diagrams of subclonal assignments for mutations at diagnosis (left) and relapse (right) between DECODE and SciClone. Widths are proportional to mutation fractions. **H**: Alluvial diagram of DECODE subclonal assignments for mutations at diagnosis and relapse. Outer bars contain mutations unique to each sample.



Supplementary Figure 9: Mutation clustering for patient AML-4. **A**: SciClone's joint clustering of mutations at diagnosis and relapse. **B-E**: Deconvolution of mutations using MOBSTER (**B**, **C**) and DECODE (**D**, **E**) for each sample at diagnosis (**B**, **D**) and relapse (**C**, **E**). **F**: Mean VAFs between diagnosis and relapse of subclones inferred from SciClone. **G**: Alluvial diagrams of subclonal assignments for mutations at diagnosis (left) and relapse (right) between DECODE and SciClone. Widths are proportional to mutation fractions. **H**: Alluvial diagram of DECODE subclonal assignments for mutations at diagnosis and relapse. Outer bars contain mutations unique to each sample.



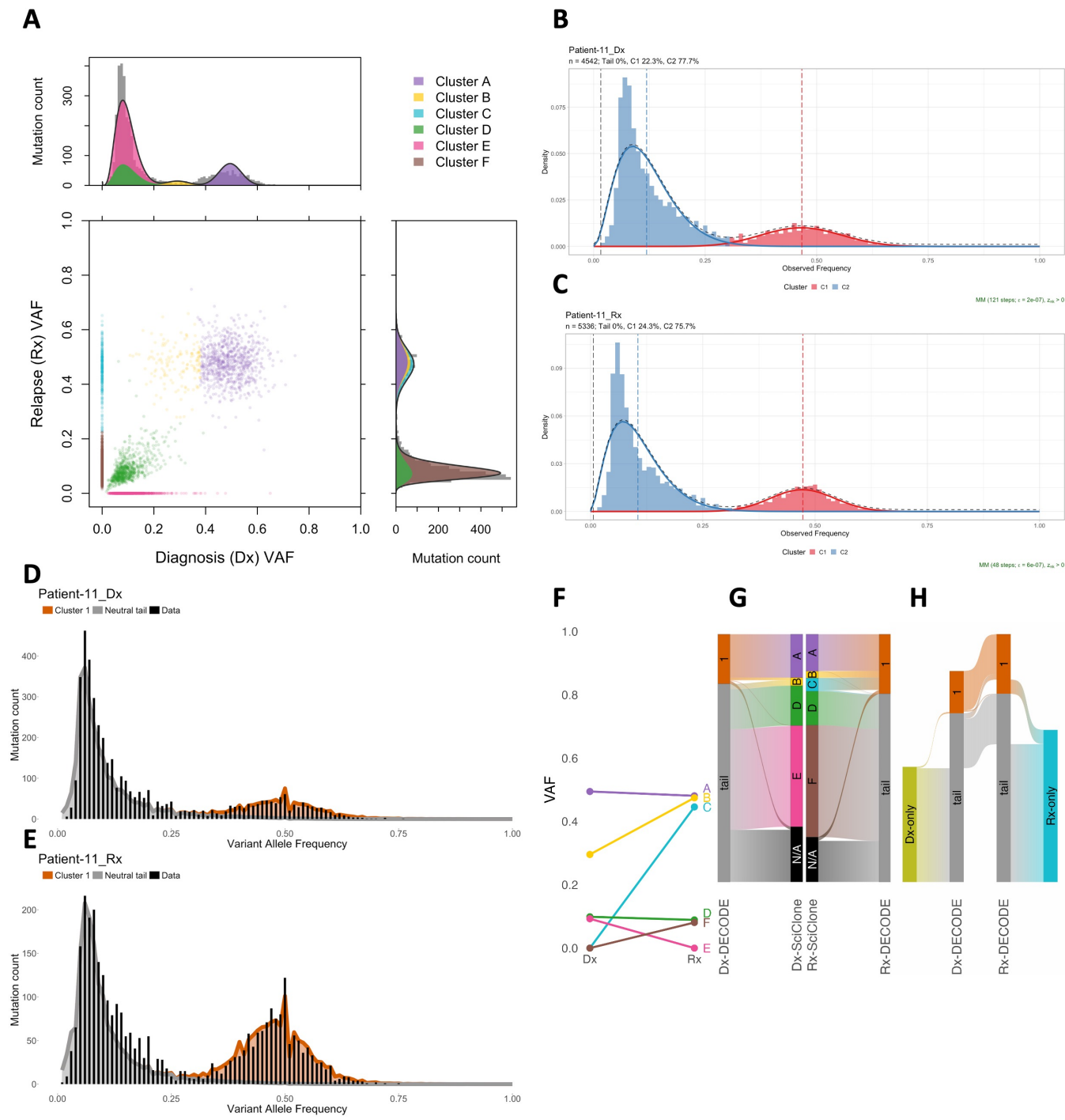
Supplementary Figure 10: Mutation clustering for patient AML-6. **A**: SciClone's joint clustering of mutations at diagnosis and relapse. **B-E**: Deconvolution of mutations using MOBSTER (**B**, **C**) and DECODE (**D**, **E**) for each sample at diagnosis (**B**, **D**) and relapse (**C**, **E**). **F**: Mean VAFs between diagnosis and relapse of subclones inferred from SciClone. **G**: Alluvial diagrams of subclonal assignments for mutations at diagnosis (left) and relapse (right) between DECODE and SciClone. Widths are proportional to mutation fractions. **H**: Alluvial diagram of DECODE subclonal assignments for mutations at diagnosis and relapse. Outer bars contain mutations unique to each sample.



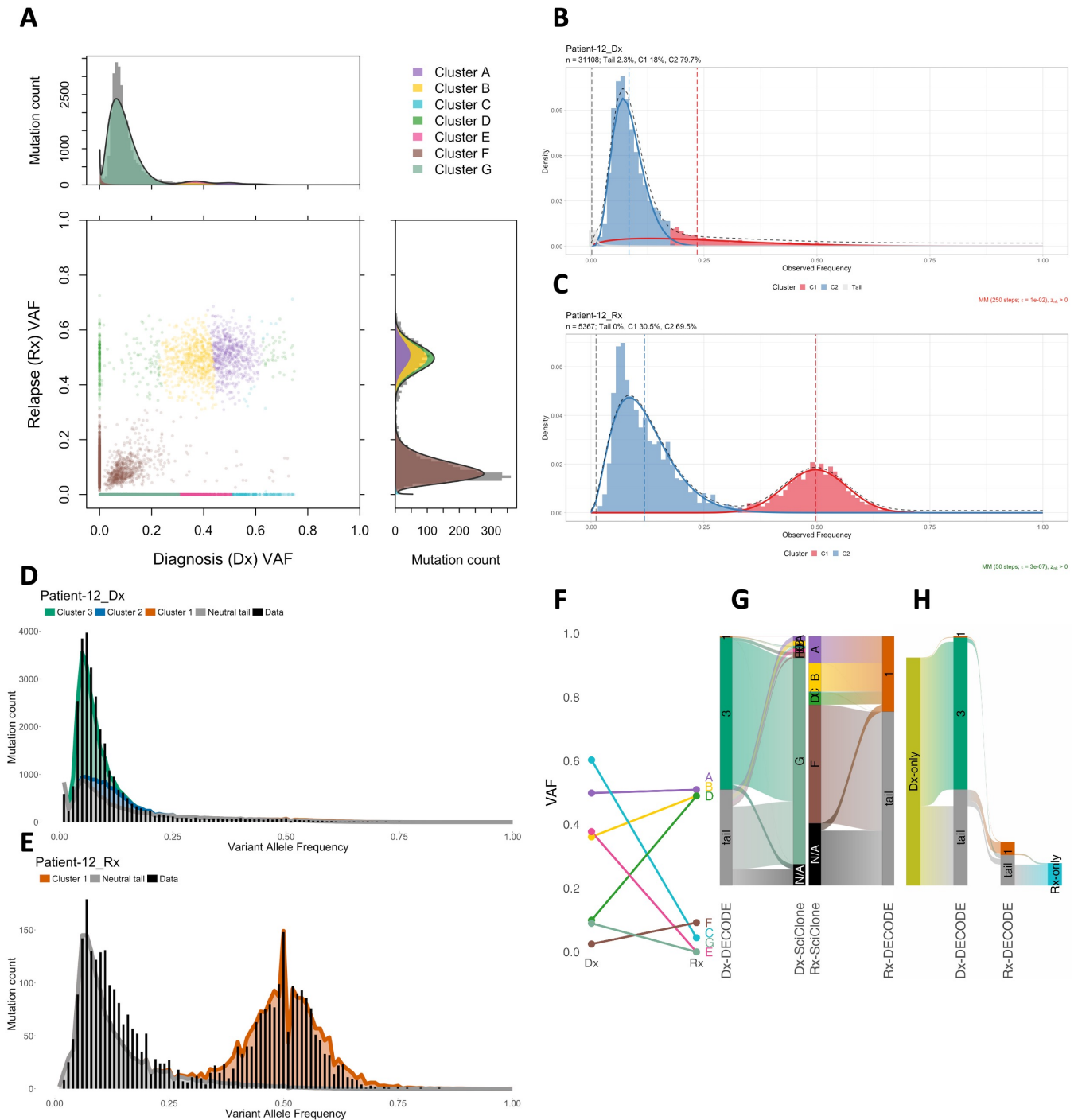
Supplementary Figure 11: Mutation clustering for patient AML-8. **A**: SciClone's joint clustering of mutations at diagnosis and relapse. **B-E**: Deconvolution of mutations using MOBSTER (**B**, **C**) and DECODE (**D**, **E**) for each sample at diagnosis (**B**, **D**) and relapse (**C**, **E**). **F**: Mean VAFs between diagnosis and relapse of subclones inferred from SciClone. **G**: Alluvial diagrams of subclonal assignments for mutations at diagnosis (left) and relapse (right) between DECODE and SciClone. Widths are proportional to mutation fractions. **H**: Alluvial diagram of DECODE subclonal assignments for mutations at diagnosis and relapse. Outer bars contain mutations unique to each sample.



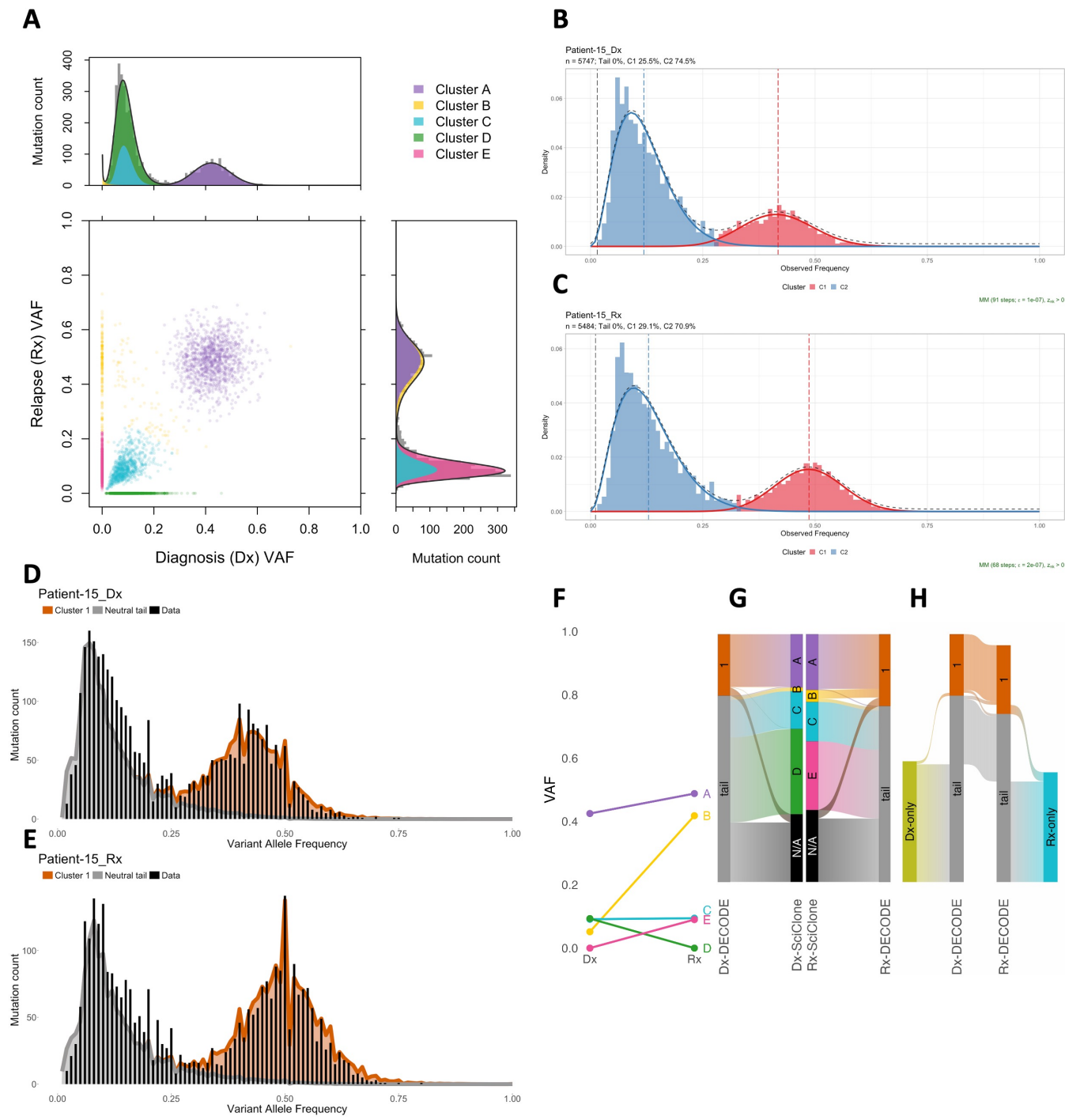
Supplementary Figure 12: Mutation clustering for patient AML-10. **A**: SciClone's joint clustering of mutations at diagnosis and relapse. **B-E**: Deconvolution of mutations using MOBSTER (**B**, **C**) and DECODE (**D**, **E**) for each sample at diagnosis (**B**, **D**) and relapse (**C**, **E**). **F**: Mean VAFs between diagnosis and relapse of subclones inferred from SciClone. **G**: Alluvial diagrams of subclonal assignments for mutations at diagnosis (left) and relapse (right) between DECODE and SciClone. Widths are proportional to mutation fractions. **H**: Alluvial diagram of DECODE subclonal assignments for mutations at diagnosis and relapse. Outer bars contain mutations unique to each sample.



Supplementary Figure 13: Mutation clustering for patient AML-11. **A**: SciClone's joint clustering of mutations at diagnosis and relapse. **B-E**: Deconvolution of mutations using MOBSTER (**B**, **C**) and DECODE (**D**, **E**) for each sample at diagnosis (**B**, **D**) and relapse (**C**, **E**). **F**: Mean VAFs between diagnosis and relapse of subclones inferred from SciClone. **G**: Alluvial diagrams of subclonal assignments for mutations at diagnosis (left) and relapse (right) between DECODE and SciClone. Widths are proportional to mutation fractions. **H**: Alluvial diagram of DECODE subclonal assignments for mutations at diagnosis and relapse. Outer bars contain mutations unique to each sample.



Supplementary Figure 14: Mutation clustering for patient AML-12. **A**: SciClone's joint clustering of mutations at diagnosis and relapse. **B-E**: Deconvolution of mutations using MOBSTER (**B**, **C**) and DECODE (**D**, **E**) for each sample at diagnosis (**B**, **D**) and relapse (**C**, **E**). **F**: Mean VAFs between diagnosis and relapse of subclones inferred from SciClone. **G**: Alluvial diagrams of subclonal assignments for mutations at diagnosis (left) and relapse (right) between DECODE and SciClone. Widths are proportional to mutation fractions. **H**: Alluvial diagram of DECODE subclonal assignments for mutations at diagnosis and relapse. Outer bars contain mutations unique to each sample.



Supplementary Figure 15: Mutation clustering for patient AML-15. **A**: SciClone's joint clustering of mutations at diagnosis and relapse. **B-E**: Deconvolution of mutations using MOBSTER (**B**, **C**) and DECODE (**D**, **E**) for each sample at diagnosis (**B**, **D**) and relapse (**C**, **E**). **F**: Mean VAFs between diagnosis and relapse of subclones inferred from SciClone. **G**: Alluvial diagrams of subclonal assignments for mutations at diagnosis (left) and relapse (right) between DECODE and SciClone. Widths are proportional to mutation fractions. **H**: Alluvial diagram of DECODE subclonal assignments for mutations at diagnosis and relapse. Outer bars contain mutations unique to each sample.